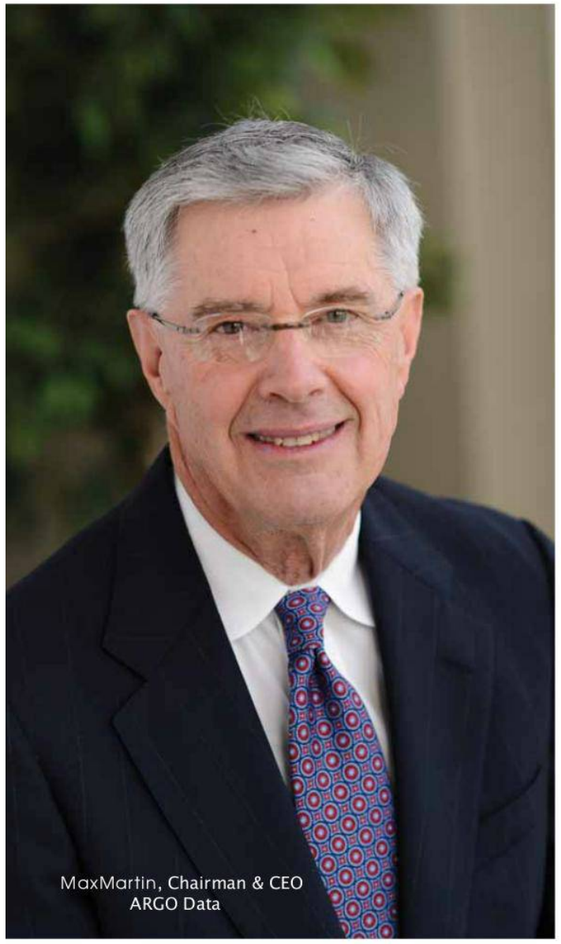
Managing Data & Improving Accuracy



MaxMartin, Chairman & CEO
ARGO Data

Working with Big Data enables organizations to solve complex problems—whether an organization is a retail operation that's open during the day with machines collecting data and performing analysis overnight, or a healthcare organization, which needs an immediate answer when a patient appears in the emergency room.

In today's environment, keeping up with the rate of data coming in and the size/amount of data needed to solve problems and achieving accurate results are key challenges.

Here are tips to manage data and improve accuracy.

Build an expert team

To effectively utilize Big Data, organizations require an expert team. An ideal team consists of:

- Computer scientists
- Statisticians
- Mathematicians

The team's background should include:

- Staff with advanced degrees, who bring knowledge on data mining and artificial intelligence in order to gain valuable insights from data.
- Employees with technical skills-both analytic and engineering to design high-performance applications to pare down and analyze the data using the latest technology.

Putting together the right team with the right mix of skills is challenging. New analytical techniques have emerged in the last 5-10 years, so it's important to hire staff with experience using these new approaches.

Identify and Measure Key Performance Indicators

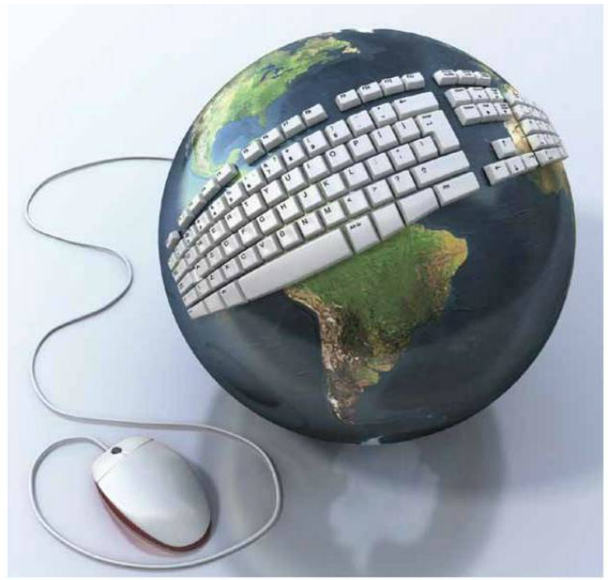
Key performance indicators allow organizations to identify top strategic and operational goals including customer satisfaction, investment and asset values, sales, and profits. These indicators provide an actionable scorecard to monitor, report on, and achieve results.

Because of the magnitude of input and output data, it's important for organizations to prioritize the top indicators and formulate multi-year plans, while at the same time identifying short-term goals. In addition, organizations need to determine how to measure performance by collaborating with internal analytics team members and subject matter experts.

Separate the noise from the data

Extracting meaningful information from raw data is a critical step in processing. Organizations have many techniques to choose from including:

- Data mining—Tools and techniques to extract information, structure, and patterns from large databases. One example is parsingand segmentingdatasuch asnames and addressesto gather important data values.
- Machine learning—Field of study that uses algorithms to automatically learn and reduce errors based on feedback.
- Natural language processing—Computerized approach to analyzing text based on a set of theories and technologies utilized by an organization.
- Predictive Modeling—Statistical models, which collect and analyzedata to predict outcomes from a complex set of input data.



Increase scale as volume increases

An effective way to manage data as the volume increases isthrough data grids and compute grids. Data grids enable users to spread the data across a set of connected machines. Compute grids take large computations and divvy them up across a number of computers.

Software systems for data grids and compute grids include:

- Hadoop
- GridGain
- Pivotal GemFire

Additional tools to increase the effectiveness of data grids and compute grids include:

- Mahout for machine learning and data mining
- Pig,Hive, and Cascading for a parallel computation framework

Large efficiencies are gained when data grids and compute grids are combined, using a MapReduce architecture. This architecture supports many computers completing analysis and calculations across a distribution (Map) or a set of connected machines. The results are collected back onto a single computer (Reduce), which does the final reconciliation.

To add flexibility, consider elastic scalability. With this approach, computers are addedor removedfrom the grid, as needed without taking the system offline. For example, an organization useseight computers Monday through Thursday to completeits processing. But on Fridays and Saturdays, the organization sees a rush of activity.To optimize the efficiency of computing resources, the organization sets up the system so that on Friday and Saturday, 20 machines handle the processing.

Utilize a multi-stage approach

A multi-stage approach enables each stage of processing to reduce the amount of data, enabling more sophisticated computations to be completed in a reasonable time frame. Through the filtering and reduction of data and the development of software, algorithms, and hardware, the data-intensive computations provide timely and meaningful analytical results.

One way to begin decreasing the volume of data is selecting particular characteristics, attributes, or fields and comparing those attributes against entries in the database. The analysisbreaks the data down into separate groups, and filters those groups. This method efficiently pares down the data and ensures computational tractability, meaning the factors that go into making a decision are kept small enough to allow the process to return an answer in a reasonable amount of time. Once the search space narrows to something more manageable, then more extensive analysisis performed.

Recognize when to use batch versus real-time processing

Organizations face different challenges when doing batch versus real-time analysis. In batch processing, analysis is typically completed at the end of the day and reports are created overnight for the next day. For example, an organization needs to forecast salesfor September and has all the sales data—for all department stores for the last 20 years—and distributes the data across a set of machines. The computersperform thecomputations and analysis for all the sales figures over several hours.

The same organization might require sales totals by week and by product line in a matter of minutes and uses real-time processing. In this scenario, it's not possible to compare results against the entire data set. Instead, the organization maintains partial results or summaries to aggregate the data for more timely responses.

Address key challenges

Keeping up with the rate of data coming in and the size/amount of data needed to solve problems and achieving accurate results are key challenges. To address these challenges, organizations require the right team and the right selection of software systems and tools. In addition, the use of a multi-stage approach and recognizing when to use batch processing and when to use real-time analysis enables organizations to effectively utilize Big Data to solve complex problems. CS