



Critical Mass Intelligence

Gaining Competitive Advantage Through Data Analysis

Contents

Introduction	2
Critical Mass Intelligence (CMI)	3
Key Elements of CMI	3
Scalability	4
Scope	4
Performance	5
Flexibility	5
Open Scale	5
Open Systems	6
Open Standards	6
Open Ecosystem	6
Economics	6
CMI and Greenplum	6
Scalability and Scope	7
Flexibility and Performance	7
Economics	8
Examples	8
Entertainment and Media	8
Financial Services	8
Telecommunications	9
Publishing	9
Other Cases	9
Conclusion	9

Introduction

Business Intelligence (BI) has become a powerful weapon in today's increasingly competitive business landscape. The insight that BI brings can significantly enhance such fundamental capabilities as strategy formulation and operational effectiveness – literally, what a company does and how well it does it. Companies that successfully exploit BI can gain greater insights into not only their internal operations but also their external environments and the interplay between the two. As a result, they can more accurately gauge and forecast the impact of past and future strategic and tactical moves from their current positions. They can more effectively allocate resources to achieve best practices in one area, optimally time a new strategic push into a second area, adeptly exit a third, and so on. Effective BI delivers the ability to tune strategies, positioning and execution to deliver lasting competitive differentiation, the key element of profit and success.

Critical Mass Intelligence (CMI)

The meaning and magnitude of BI are different for every organization, from a sole proprietor dissecting a handful of inventory spreadsheets, to teams of analysts scrutinizing petabytes of data at one of the world's largest corporations. The results are almost as diverse as the approaches. The "intelligence gap" separates leaders that have established BI excellence that exceeds industry norms from followers that have not. What is it about successful BI implementations, then, that separates leaders from followers? What is it that allows some organizations to use BI as a competitive weapon, and keeps others several steps behind? What does it take for organizations to jump the intelligence gap and build real, sustainable competitive differentiation?

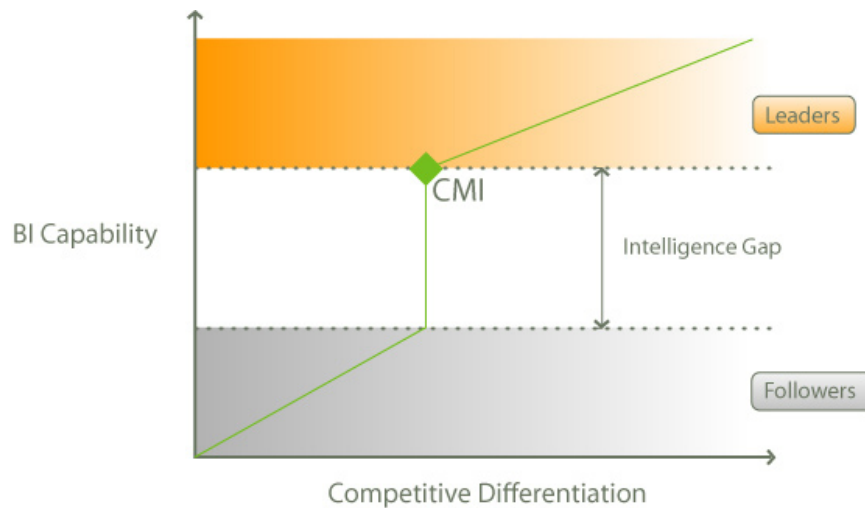


Figure 1: CMI, Competitive Differentiation, and the Business Intelligence Gap

Greenplum calls this concept **Critical Mass Intelligence**. Critical Mass Intelligence (CMI) is the state an organization reaches when it attains the ability to achieve insights of sufficient magnitude, scope and flexibility to deliver superior operational effectiveness and tune its strategy to the extent that it can produce lasting competitive differentiation in its market.

Key Elements of CMI

In most industries, only a minority of companies have achieved critical mass with their BI capabilities. Without a solid understanding of the key requirements of CMI, managers and practitioners alike tend to gravitate to products that worked well in other (non-BI) areas of the IT infrastructure, in hopes that the old tools will work as well in BI as they do with other applications. No amount of hope can create breakthrough results, however. Developing a firm grasp on the components of BI that deliver results is a first step in reaching CMI.

Scalability

Data is the raw material of BI. All other things remaining equal, access to more data is better than access to less. To reach CMI, a firm must be able to access and analyze the data that matters most to the success of the organization. In today's information-driven economy, a single data set can occupy terabytes or even petabytes of storage capacity. Companies whose BI infrastructures are not architected to function at large scale are often forced to choose between ignoring entire data sets altogether, or operating on aggregates and partial data sets. Relying on anything other than the entire population in a given data set reduces the accuracy of information (and the confidence levels of that information) that feeds decision-making, which can result in suboptimal or incorrect conclusions, and limited ability to realize valuable insights.

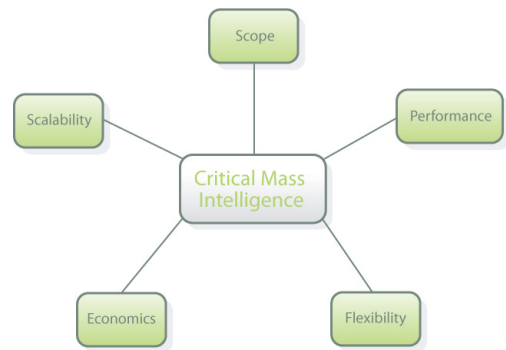


Figure 2: The Five Key Elements of CMI

The primary onus of supporting large scale BI is squarely on the database management system. What makes this particularly difficult is that traditional databases can rarely adequately support more than a few hundred gigabytes of data for BI workloads. Because most companies use traditional database systems to support transaction-oriented (non-BI) operational applications, most try to use the same products to support their BI applications, at least until the scalability and performance limitations of those products make supporting them untenable. A widely accepted BI industry rule of thumb is that traditional, transaction-oriented database systems start to break down in their ability to support BI workloads at around one terabyte of data. Firms that attempt to scale traditional databases beyond that level rarely succeed. Despite spending very large sums of money on additional hardware, software, and professional services for query rewriting and database tuning, these organizations generally fail to reach their goals.

Scope

For most companies, a single data source does not represent enough information to achieve material insights. Multiple data sources of significant volume are generally required to achieve CMI. Although reaching critical mass generally does not require integration of every data source at the company's disposal into a single, monolithic BI system, more sources can often yield more, and more substantial, insights.

Each data source may have a radically different originating application, format, and schema. The BI infrastructure must be able to accommodate the structure and relationships that bind multiple data sources together. Many database systems do not support arbitrary database schemas, let alone complex, multi-join queries against arbitrary schemas.

Moreover, some products, most notably column-oriented databases, require DBAs to configure and execute complex pre-aggregation and projection routines. These routines take a snapshot of the customer's workload and data, and decide how to organize the data on disk. The failed assumption in this model is that the given workload snapshot is both representative of the current situation and static over time. In practice, workloads and schemas are dynamic.

With systems that support significantly limited ability to handle the complexity and dynamism of today's BI workloads, schemas and queries, i.e. "scope", the comprehensiveness of data access suffers. The ability of analysts to achieve insight of sufficient magnitude and flexibility suffer along with it, and CMI remains unattained.

Performance

In addition to simply storing large quantities of heterogeneous data sets, BI systems must also provide high performance data processing to achieve CMI. The concept of “data processing” as it relates to CMI can take several forms that together compose the workload of the BI database system, including responding to queries, modifying existing data or schema, adding new data to the analysis set, and other data and database operations.

Query speed is one of the most important components of the performance of a database management system. When an analyst waits too long for a response to a query, he tends to lose his train of thought. With the user's train of thought go the odds of generating significant insight. True CMI requires that BI systems deliver response times that are far closer to what technologies like Google search provides than what traditional database systems can deliver. In most cases, traditional database technologies are incapable of delivering such performance on data volumes near or beyond one terabyte.

The need to quickly modify existing data and schema is also important, as is the ability to quickly and frequently load vast quantities of new data quickly. Some database systems, and column-store databases in particular, are ill-equipped to support rapid data and schema changes. Modifications require specialized skills and can take hours or days to complete. These database systems also load new data very slowly because they do not execute loads in parallel. Furthermore, architectural constraints of column-style database systems make it difficult or impossible for users to query the system while new data is being loaded into the system. In today's business environment, the pace of business dictates that loading operations occur daily, hourly, or on an even more frequent basis. These limitations can severely limit the window of time during which generating insights is possible, and therefore reduce firms' chances of turning the corner with their BI initiatives.

Another key facet of performance is reliability. If analysts cannot depend on BI system, they cannot reasonably expect to reliably uncover new insights, and CMI is therefore unattainable. BI systems must therefore deliver sophisticated fail-over, recovery, backup and restore facilities to protect customers from potential disaster and maintain availability. Beyond simply providing RAID for redundancy, systems should employ inter-host redundancy and facilities for rapid recovery from a variety of types of failures with no loss or corruption of data.

Flexibility

CMI requires that the BI system be ready and able to change to adapt to altered circumstances. Flexibility of the BI infrastructure along the axes of scale, systems, interfaces, and vendor ecosystem are critical to reaching CMI.

Open Scale

The system must be flexible in its ability to scale to support rapidly growing data volumes, and therefore must possess the capability to easily scale beyond its initial capacity. A CMI-compatible BI infrastructure must include a database system that allows for inexpensive, modular scalability. Horizontal scalability, i.e. the ability to add inexpensive nodes to an existing system to augment capacity and performance, is imperative. Traditional database systems that were designed for transaction-oriented workloads do not support horizontal scalability. Adding capacity and performance to an existing system with these products often requires a wholesale replacement of a monolithic and expensive central server. An expensive and time consuming proposition, this phenomenon is known in IT circles as the “forklift upgrade,” and is not compatible with the principles of CMI.

Open Systems

CMI requires open hardware support, which allows customers to select their systems vendor and server platform of choice. Open hardware is general-purpose hardware available off-the-shelf from a wide variety of vendors (including systems from top-tier hardware vendors as well as white box vendors) that allows customers to benefit from the capabilities, convenience and low costs of commodity hardware. Many BI systems rely on proprietary, non-standard, “closed” hardware. Proprietary, hardware-based interconnects and field programmable gate arrays (FPGAs) are examples of closed hardware. These products require that any upgrades or replacements be purchased from the original hardware vendor. Organizations that select proprietary, closed systems are effectively betting that a single vendor will innovate ahead of the curve in not only software, but hardware and software simultaneously, and that the vendor’s innovation rate and quality levels will equal or exceed those of an industry full of pure-play hardware vendors.

Open Standards

Support for open standards is another key component of flexibility. Support for hardware standards such as x86 processors, SATA and SAS disk technology, and gigabit Ethernet networking help to future-proof the BI system and protect against obsolescence on the hardware side of things. Also important are support for full ANSI-SQL, ODBC and JDBC. These standards ensure compatibility with the widest possible variety of BI tools and applications. Many proprietary BI vendors forego support for open standards in favor of proprietary hardware. Others fail to fully support ANSI-SQL standards. This not only forces customers to learn radically new ways of interacting with the data management platform, but also forces them to decide between retooling their existing applications and abandoning their applications altogether in favor of entirely new ones. History has shown that betting against standards can be a risky proposition.

Open Ecosystem

Organizations seeking CMI need to be able to freely choose the components of the overall BI solution, whether it’s database software, tools, applications, or professional services. Some BI vendors combine all of these elements, in attempts to lock customers into a single vendor for the entire BI stack. Rarely do companies find that a single vendor can deliver all of the technologies and services required to attain CMI.

Economics

One final element required to reach CMI is cost effectiveness. For decades, only the largest and richest corporations could afford large-scale BI. For many of the lucky few that could afford it, BI was transformational. Adoption marked an inflection point in sustainable competitive differentiation and value creation. Even in today’s marketplace, traditional BI solutions regularly cost upwards of \$800,000 per terabyte of capacity. Very few companies can consider that magnitude of investment to manage an amount of data that is modest enough to fit on today’s consumer desktop computers. As a result of not only the inadequacy of traditional solutions but also their extreme costs, few companies have reached CMI.

CMI and Greenplum

Greenplum was founded to allow companies to achieve and surpass CMI regardless of whether or not they wield massive budgets on par with those of the Fortune 100. Greenplum’s approach to achieving CMI is to deliver scalable, high performance, flexible Data Processing Networks (DPNs) to our customers. Greenplum DPNs comprise clusters of powerful, commodity servers running Greenplum Database, and they represent the heart of a CMI-compliant BI infrastructure.

Greenplum DPNs integrate seamlessly with other key elements of the infrastructure, including front-end BI tools, applications, and Data Integration (DI) technologies. The cornerstone of any Greenplum-powered DPN, though, is the database itself. A truly breakthrough software product that began with an open source core, Greenplum Database is a Massively Parallel Processing (MPP) database architecture optimized for deployment on powerful, low-cost, general-purpose systems. The result is a product that is uniquely capable of propelling BI initiatives to, and beyond, CMI.

Scalability and Scope

Greenplum Database was purpose-built for high performance BI, and data volumes ranging from terabytes to petabytes. The product supports arbitrary schemas and sophisticated queries, which it executes with blinding speed without application retooling or database tuning. Complex joins are supported out of the box. These and other capabilities make Greenplum Database an excellent choice for supporting any BI or data warehousing implementation, from DPNs focused on just a handful of topics or data sources, to enterprise-wide data warehouses.

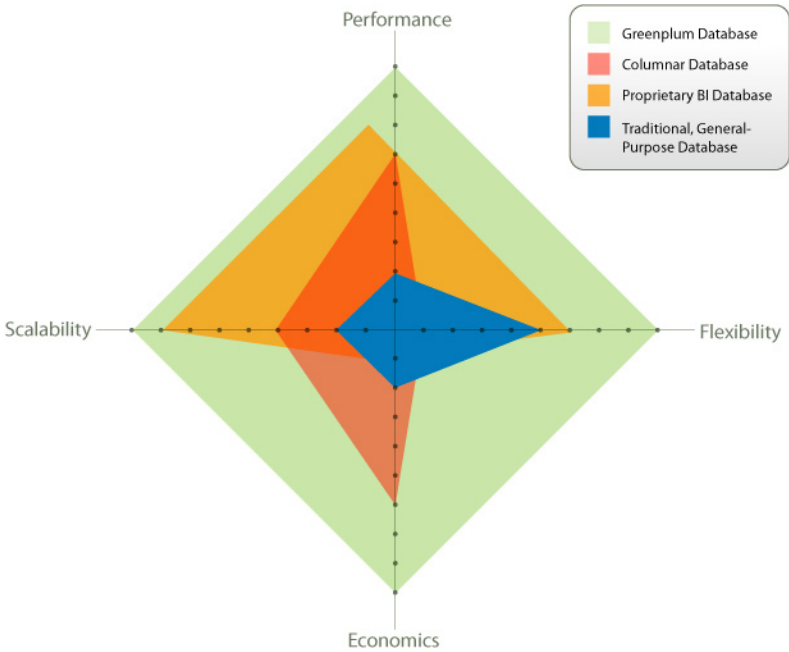


Figure 3: Measuring Database Products On the Key Elements of CMI

Flexibility and Performance

Greenplum-powered DPNs are modularly scalable, and allow customers to scale horizontally, in cost effective, granular increments. Greenplum Database’s Dynamic Provisioning technology allows customers to easily and inexpensively achieve greater capacity and performance by simply adding commodity servers to the system. Because the product’s performance scales linearly with then number of servers, implementations involving hundreds of terabytes on a single Greenplum DPN are commonplace. That said, the technology is also well suited to implementations in the single-digit terabytes.

Greenplum Database delivers query performance that is typically between 10 and 100 times faster than traditional products. It can also load 6 to 10 times faster than any other product on the market, at a rate of over 4 terabytes per hour in production systems.

Greenplum Database also supports open hardware, so organizations deploy it on the general-purpose servers of their own choosing, purchased from the hardware vendors with which they are most comfortable. There is no proprietary hardware involved in the solution, so vendor lock-in is a non-issue. Greenplum's support for open standards and interfaces and certification with leading BI, DI, and vertical-specific products make it easy for customers to use Greenplum Database with their existing applications.

Economics

Several aspects of Greenplum Database make it significantly less expensive to adopt, scale and maintain over time, relative to traditional database solutions and proprietary, black-box analytic appliances. Greenplum's open source heritage and compatibility with commodity, general-purpose hardware reduce acquisition costs. Out-of-the-box parallelism delivers performance without expensive database tuning. Adherence to open standards for not only server technology but also programming interfaces like ODBC and JDBC and syntax standards such as ANSI-SQL reduces the costs of education and maintenance. Dynamic Provisioning allows for very inexpensive scalability into the petabytes. The result of Greenplum's approach is a product that offers not only superior capabilities, but also order-of-magnitude more attractive economics.

Examples

There is a growing list of organizations that have deployed Greenplum Database and achieved or surpassed CMI. These companies represent a wide variety of sizes, business models, and aspirations, from telcos, to financial services, to transportation, to Internet and beyond. Each of them began with the desire to use technology and data as a competitive weapon. Their attainment of CMI allows them to achieve insights – true “business intelligence” – that few competitors in their respective industries can match.

Entertainment and Media

A large, U.S.-based entertainment and media company deployed Greenplum Database to enable the monetization of one of its largest web initiatives. Initially deployed in a single DPN, the system's primary fact table contains over 900 billion rows, and the overall system supports 400 terabytes of usable database capacity. The customer loads roughly 10 billion records each day, with loading speeds in excess of 4.5 terabytes per hour. The customer uses industry standard DI and BI tools as part of the DPN, and has an analytical capability that exceeds that of any other competitor in its category. The solution was an order of magnitude less expensive than the proprietary BI solution that it evaluated, and is allowing the company to achieve a level of advertisement targeting that is unsurpassed in the industry.

Financial Services

In the financial services industry, the explosive rise in algorithmic trading is driving not only the aggregate frequency and number of trades, but also the number of messages required to complete trades, to increase, resulting in exponential data growth. A handful of America's largest financial exchanges have selected Greenplum Database to store and analyze every trade executed on their respective indexes. One institution in particular had tried and failed for two years to configure and tune a traditional database to handle tens of terabytes of capacity for BI queries. Persistent problems with query performance, loading performance, system stability and cost led the company to evaluate and select Greenplum Database. The resulting system will allow growth into the petabytes and CMI-worthy performance and flexibility for a fraction of the cost, and allow the company to operate with freedom from the risks and constraints of using traditional database management systems for BI.

Telecommunications

A telecommunications giant in India purchased a 40-terabyte system from Greenplum to improve its performance with compliance data. The solution reduced the time required to retrieve detailed call records by 80%, improved loading speeds from two hours to ten minutes, and significantly mitigated the risk of falling out of compliance for call detail requests. The company recently took advantage of Greenplum Database's cost effective, horizontal scalability when it decided to expand the system by 200% to encompass additional data sources for analyzing billing and overall operational execution. As its competitors struggle to cope with the rapid data growth that is accompanying today's telecommunications market in India and Asia in general, this company is excelling, with the confidence that the capacity, performance and cost of its BI systems are well under control.

Publishing

A technical book publisher uses Greenplum Database to identify and predict trends among potential buyers of its books. It uses this information to optimize its selection of topics for upcoming books and articles. To do this, the company stores and analyzes every job posting, every blog post on the Internet, and several other large-volume data sources, looking for trends. The system reduced the company's most time-consuming and complex query from ten hours to six minutes. With Greenplum, the company's analysts can work with a BI system that delivers answers quickly enough to keep pace with the analysts. The level of insight that the system provides helps the company maintain and extend its leadership position in the category.

Other Cases

Several of the world's largest financial services use Greenplum Database to track and analyze a wide variety of financial data. Internet companies use the product to identify fraud, track online brand abuse, quantify grey market e-commerce, optimize online advertising campaigns, accurately target customers, and track and predict customer behavior. Leading airlines use Greenplum Database to optimize routes and fare prices.

Conclusion

Reaching and exceeding Critical Mass Intelligence requires a BI infrastructure that is as cost effective as it is scalable, flexible, and fast. A growing number of organizations are deploying Greenplum Database – the world's most scalable, flexible, cost effective high performance database – to reach Critical Mass Intelligence. They are using data as a competitive weapon, and are controlling their data rather than being controlled by it. Greenplum Database is the breakthrough product that is making CMI – and the sustainable competitive differentiation that often accompanies it – possible for companies around the globe.