THE "CRITICAL MASS" CONCEPT

John F. Duffy
ICI America Inc.

Introduction

The 'Critical Mass' concept is simply a different way of looking at traditional data in this, the 'Data Base Era'. It stresses the user's view of the data rather than a data processing view. The concept, if properly understood and applied, can greatly simplify the task and improve the results of data management.

The concept of "Critical Mass" can be particularly important in the development and implementation of a research data management strategy. Research applications, data, data collection, maintenance and reporting differ significantly from traditional business applications. This paper will describe the concept and show how it might be applied in a research environment.

Brief History

The evolution of data processing might be characterized as follows:

Phase I - Batch systems with a heavy emphasis on accounting applications. Expensive hardware with little or no software support. The programmer was all important.

Phase II - Still heavily batch oriented, but now the application support was expanded to manufacturing, inventory, distribution, order entry, etc. The hardware was still expensive and software limited, but now the analyst became more important.

Phase III - Shortly after the introduction of the IBM 360 line the "Critical Mass" concept was born. The 360 was still expensive but a better price performer than its predecessors. It provided systems and applications software and promised on-line and data base support. The computer was becoming more and more a management tool. The 'Critical Mass' concept grew out of a need on the part of management to gain better control of their data, and better utilize an expensive resource. The vendor was also pushing the 'Critical Mass' concept in an effort to make the 360 a better on-line processor. The original 'Critical Mass' concept stressed the elimination of data redundancy in order to better utilize disk space and permit more efficient processing.

The early days of data processing can be characterized as a series of the mechanical solutions or the automation of applications which already worked manually. The inputs and outputs were easily defined and the anticipated results readily calculated. The computer simply produced these well defined results faster and cheaper.

Today's Environment

The first 3 phases of data processing are often called the days of feeds and speeds, because the emphasis was on the machine's ability to process batch jobs.

Today and in the future the emphasis will be toward on-line systems with heavy user involvement (see Figure 1). The cost of maintaining data processing support has swung from hardware to people (see Figure 2). Productivity is our goal. Management is constantly looking for opportunities to substitute relatively inexpensive hardware to reduce or eliminate manual effort.

Business vs. Research Applications

The typical business application can be characterized as follows:

- Repetitive in nature (i.e. executed on a schedule - daily, weekly, monthly, etc.)
- Relatively long payback period
- Consistent and identifiable data relationships and processing requirements.

Critical Mass - a definition

Much has been written about the subject, but it is difficult to find a good general definition. For the purpose of our discussion, I would define the Critical Mass as:

"That collection and organization of data relative to a particular problem, task, project, application, etc., which accurately defines and maintains valid data relationship in support of consistently sound results."

Research Data and the Critical Mass

Most research data can be represented as a series of events with each event made up of one or more observations about a particular subject, process, or phenomena. The research data for any subject, process or phenomena will be composed of one or more events, each of which may or may not be repeated over time (see Figure 3).

The general definition of the Critical Mass, as stated earlier, refers to the total collection of data about a particular project, task, etc. To adapt the concept to research data it should be aimed at the event. The 'Critical Mass' concept as applied to the event is a procedure which I call:

"INSULATING THE EVENT".

Insulating the Event can be defined as:

Capturing an event with its set of observations and then storing with the event a copy of any related data elements which (1) are necessary to accurately represent the event data (2) if not copied and carried with the event could possibly change over the duration of the research project thereby destroying the ability to easily and accurately represent the event.
This definition can best be explained by using a simple example (see Figure 4). Assume a file of physicians where the root segment contains basic identification information, most of which is static (example: Name, address, specialty rating, year graduated medical school, etc.). Each time our salesmen call on the doctor we record a call event. Call information includes date, products detailed, samples, etc. When we mail literature to the physician we record a mail event with information similar to the call event. This file design appears sound as viewed from the traditional 'Critical Mass' concept. However, if we apply the concept of 'Insulating the Event' we find the file design to be lacking. If the doctor changes his specialty (GP to IM, etc.), or if the salesman changes the doctor's rating, it would become virtually impossible to accurately represent prior calls relative to specialty and rating.

In the case just presented, insulating would require that the doctor's rating and specialty be copied and carried with the Call or Mail event, thereby insuring accurate analysis and representation.

In our example, which is a business application, this problem would eventually be detected and corrected. However, in research where similar problems exist, we do not always have the luxury of time and repetition. And so, Insulation can provide a fast and relatively inexpensive solution to this type of problem.

Insulating the Event accomplishes several things, all of which are in line with the typical research application and supported by the long term trend in data processing.

Insulation does cause increased data redundancy. However, with the reduced cost of data storage the redundancy is more than offset by the increased facility for analysis. Also, research projects are generally small in relation to business applications.

Insulation reduces the need for significant front-end planning on research projects which often have a short life span since questionable data relationships are firmly established through data redundancy.

Insulation provides the researcher with a simplified view of the data. In this era of online access and sophisticated data management and retrieval tools the researcher can carry out his or her inductive or deductive search without having to mentally construct complex logical relationships or modify his logic to compensate for a change which now might effect some prior events.

In most business applications results can be anticipated and therefore it is easier to identify logic errors and correct them. However, in research results are difficult to anticipate and so errors are difficult to detect and correct. If the data base loses creditability the project is worthless. The cost of a little redundancy (INSULATION) is a small price to pay for increased creditability and productivity. In research 'Critical Mass' means 'Insulating the Event'.

Figure 1

**Growth of online information systems**

<table>
<thead>
<tr>
<th>$ billions</th>
<th>Value of installed products</th>
</tr>
</thead>
<tbody>
<tr>
<td>1974</td>
<td>22.6</td>
</tr>
<tr>
<td>1980</td>
<td>36.1</td>
</tr>
</tbody>
</table>

Source: Quantum Science Corp.
Figure 2
Application development costs

Percent DP expenditures

<table>
<thead>
<tr>
<th>Year</th>
<th>Equipment</th>
<th>Personnel</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>1955</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1965</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1975</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1985</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total DP expenditures

Figure 3

Project

Subjects

Events

Observation

Events

Observation

Events

Observation

Events

Observation
Figure 4

Root

Physician I.D. Info.

Event

Call

Mail

Date

Products

Samples