

The Data Czar

Introduction

As the Information age matures a variety of problems regarding the nature of the efficiency, effectiveness, and productivity of information and data systems have become apparent. Because analysis of the Data Systems are often initially focused on the result of problems, the systems are often labeled as inefficient, ineffective, and unproductive without due diligence and cause analysis. Clearly, there are instances when the problems associated with a given data flow model are solved by the addition of increased processing power or network bandwidth. However, the majority of Data System problems can be significantly, if not exclusively, alleviated by an intelligent technical and logistical analysis of data flow from the perspective of removing bottlenecks associated with data flow to and from the Clients, File Servers, and Disk Farms.

Any company that is serious about maximizing output will, or should, employ a data management specialist. This person's role is more logistical and political than technical, and as such the person can be referred to as the Data Czar. The key to implementing a successful data strategy is to recognize the perspective, limitations, and workflows of the actual users. The knowledgebase present in the user community must be incorporated into the data flow.

Understanding the data flow to gain efficiencies.

The Data Czar responsibilities encompass a broad spectrum of the data flow model to facilitate and enhance the ability to allocate how data needs to be managed. The management of the data is vital to increase the efficiency and productivity of the data system.

- Optimal Data Flow – Data is often transferred from one application to the next application, or from one department to the next department, where the output of one process becomes the input of another process.
- Optimal Use Of Duplicate data - In many cases, duplicate data wastes space. However it is precisely the intelligent use of duplicate data throughout the system that can enhance input and output characteristics.
- Incorporating Schedule and Work-Flow - The intelligent understanding of the workflow schedule is critical to understanding how to get the right data in the right place at the right time. Any fully successful data management solution must understand the production schedules and use actual data to facilitate the creation of an accurate schedule.
- Data Control – Data is a corporate asset; therefore, the control of the data is what a data management system is all about.
- Working Data Versus Final Data - Versioning and Publishing is a vital part to any construction oriented data system; however, companies often confuse Asset Management with Data Management.

| | |
|------------------|--|
| Asset Management | Tracks and manages revisions of individual assets (example a character), or a collection of individual assets that comprise a higher level asset (example a frame) |
| Data Management | Manages the quantity, location, and performance requirements of data to support the production pipeline in the most efficient manner. It also tracks and manages actual physical hardware resources. Data management, typically is unconcerned with the context or metadata. |

The data system must optimize both the management of data and assets.

- Primary Data and Value Add Data - The intelligent understanding of primary data and value add data is key to maximizing efficiency in the data system. Primary data often consists of the original source data, and intermediate data resulting from an automated process applied to the source data. Primary data is often re-creatable, but at some cost. Value add data refers to data with some human value add, which is typically expensive to create and typically not 100% re-creatable. Value-add data often requires already scarce human resources to create.

The Data Czar should assess where a company is in their evolution and approach to data management, then guides them through the best practices to improve efficiency.

Without a thorough understanding of the data system, companies often struggle with some of the following:

- Can you account for all the data on your file systems?
 - Who put it there, in what time frame?
 - Who is currently responsible for it or using it?
 - Why are we running out of space? What can we get rid of?
 - Can it be deleted? Who should we talk to within the organization?
 - Managing less data is an effective data management process!
- How does data management process help validate, scrub, or QC data on the file systems?
- How is the network of complex relationships among the data sets incorporated into the data management policies? i.e. this file required, for this build, etc? or is it common to all builds?
- Can data be visualized by project, product, design, user, department, etc?
 - Can this information be viewed by project leaders on a regular, real time basis? For example, if we need to bring a new project online, is there enough space? Can space be taken away from another project? If this information is being extracted via file system, user, & group permissions... what if they are not reflective of true owner, or user of the data... what if file system permissions are fairly wide open by department?
- Data Forensics -> who deleted, moved, etc what data, when?
- How is the relationship of data to production schedules, objectives, etc. incorporated into the data management system / migration policies? For example, do we demote all data associated with this project because the project is on hold? How often / frequent are these policies updated to reflect changes in schedules & objectives?
- Can all the data for an entire project (or component of a project) be archived or retrieved?
- Can the knowledgebase for data management decisions be extended to include important information resident in the user community?
- How can discipline and control be introduced into corporate data assets? (how open are file system permissions)
- Can business logic, workflow, and policies be incorporated into file systems? How does this align with my data management policies? How labor intensive is this process? (data classification, to set data management policies) If labor required, how well will this scale?
- How can data management be proactive to user requirements, as opposed to reactive?
- How is all of the above information obtained, which is required to effectively manage data flow, and control costs?
- Can data management policies be differentiated for "working data", quasi-temp data, and "published data assets", or is the working area the spot for "user's garbage"
- Can the organization of my data be automated.... so we don't wind up with a directory called "Joes_Junk" or "Joe_Final"
- If file system is blown away, can we recover / recall the most important project first?

Underlying inefficiencies in data management often manifest in some of the following symptoms:

- Constantly running out of disk space and don't understand why
- Routine "emergency storage purchases"
- Difficulty finding data
- Difficulty determining quality of data
- Network seems slow (often a result of poor data distribution, and many processors converging on a given partition)
- Difficulty accounting for all data on your system (why is it there, what project does it belong to, can it be removed, who should we talk to or who is responsible for the data?)
- Output not completing in a timely fashion (renders, compiles, interpretations, etc.)
- People constantly staying late, fighting fires & daily processing, to manage output.
- Companies often in a data pipeline crisis are too busy "sticking their fingers in the dike" to address the "root cause" to stop the vicious cycle and help themselves
- User perception "why do we have an IT group, if we need to worry about data... that's their problem" true, but even if you stopped generating data today... IT will own the ongoing bill for data already generated for many years to come... which then becomes a company / finance problem.
- Just give us the space, we will manage it... wind up with blocks of data scattered about their playroom. No one addressing efficiency of data across departments and organization as a whole... only choice becomes to continue to react and buy more storage.