KEYPOINT INTELLIGENCE

SERVICE AREA:

Business Development Strategies

ANALYSIS

BIG DATA'S ROLE IN THE PRINTING INDUSTRY

WHERE WE ARE AND WHERE WE NEED TO GO

JULY 2022

contents

Table of Contents

Table of Figures

Introduction

Big data involves enormous data sets from various sources that need management and access speeds far exceeding traditional relational database capabilities. Big data can drive predictive analysis for trends and opportunities that cannot be ascertained via structured relational database queries. The printing industry gathers large amounts of job, customer, and processing data now that today's print service providers (PSPs) handle most jobs digitally. Such data could be treated as big data with analytical tools to address preventative services, process improvements, staff effectiveness, and other areas.

When they hear the term big data, most people probably think of pop-up ads that appear based on browsing history. While this is a prime example, big data has more sources than social media and applications beyond online upselling, including the potential to streamline printing operations. Oxford Language defines the term as follows:

> *big da·ta*
>
> *noun COMPUTING*
>
> *extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.*
>
> *"much IT investment is going towards managing and maintaining big data"*

While accurate, this definition does not capture the difference between big data and very large databases using structured query language (SQL) coding. The differences between these two concepts are significant because they change how providers develop and maintain big data software solutions. Thus, big data solution programmers require a different level of understanding, education, and skills than traditional database developers.

Understanding Big Data

The big data term first became popularized at the turn of the century as innovative companies began using cloud-based servers and capturing vast amounts of data.

- 1998: Google began storing search data online.
- 1999: Salesforce.com began collecting data on sales and sales leads.
- 2002: Amazon's web-based retail services began gathering vast sales data.
- 2004: Facebook began capturing and storing personal information in the cloud.
- 2005: YouTube began adding vast amounts of video into the cloud.
- 2006 & 2010: Amazon Web Services (AWS) and Microsoft Azure launched, brining cloud-based computing to companies on a global basis.

New social media networks and other services continue to add to the mix, but an enormous amount of data has been captured and stored online in the past decade. People began to realize the potential value of this data if there were ways to make sense of it. The challenge, however, was that this data was not stored in a single database with a well-defined schema—and there was a ton of it!

By the early 2000s, many researchers, mathematicians, and analysts were attempting to comprehend and express the unique characteristics of big data so that it could be analyzed. One industry analyst, Doug Laney (currently Innovation Fellow of Data & Analytics Strategy at West Monroe) came up with today's commonly accepted big data description. Laney is an author, and his definition caught on because it was correct and catchy. It also condemned those who study big data to suffer through alliteration.

Laney differentiated big data from large datasets with what he termed the three V's: Volume, Velocity, and Variety.

Volume

The volume of data that is generated and stored in the cloud is almost incomprehensible. According to an often-quoted statistic from 2020, Facebook generated 4 petabytes (4 million gigabytes) of data per day and stored it in what's called the Hive. In 2020, the Hive contained 300 petabytes of data. I was unable to find more recent statistics, but the point is that Facebook alone generates massive quantities of data that would cripple traditional data analysis tools. This is a critical insight because storing big data is only part of the problem. The more significant challenge is to identify relevant data within these enormous data sets and to make sense of it.

Velocity

An endless torrent of data is streaming into what analysts call data lakes. More importantly, though, software solutions and users often need to make near-real-time decisions based on this data. If there's a bottleneck in the production line, managers want to know about it immediately. If an account has a press that is showing signs of impending failure, service technicians need quick updates. This need brings a specific set of challenges and requires a different approach than crunching end-of-quarter sales results to determine trends.

Variety

Big data often contains information from a variety of sources in a variety of formats. These might include PDF files, videos, XML documents, structured text files, Internet of Things (IoT) data streams, or other sources. Consider the data surrounding a print job:

- Customer information
- Job ticket information
- emails

- Estimates, pricing, and accounting data
- Job logs
- Job Definition Format (JDF) files
- Real-time barcode scans at workflow stations

This data comes in three different varieties:

- **Structured data** has a predefined schema to organize the data. Job logs are good examples of structured data with columns for job number, job name, quantity, etc.
- **Unstructured data** does not have a schema. Consider an e-mail message or a video. There can be valuable information inside these items, but it is not structured in a way that traditional relational databases can understand.
- **Partially or semi-structured data** has data delimiters but is not laid out in a database-friendly schema. Extensible Markup Language (XML) and JDF files are prime examples of partially structured data. These files contain tags that define data elements but are not formatted in a database-friendly schema.

Combining these data types and source varieties only compounds the problem. Structured data might be structured differently when arriving from different sources, which can cause issues for those who are trying to make sense of it.

Of course, many more minds have considered the dig data subject in the past 20 years and have augmented the initial three Vs. They even extended the alliteration to include another V—Variability. The initial three Vs discussed a variety of data sources and formats, but they did not consider variability over time. As the world changes, the types and meanings of the data we use also evolve. This adds a temporal complexity to big data.

In addition, IBM data scientists have added another three Vs into the mix. While I think they are valid points that are worth mentioning, I do not consider them unique to big data.

1. Veracity: In the eighties, we used the phrase "garbage in – garbage out." Not much has changed; data quality is still critical. Analysis based on tainted and inaccurate source data drives unreliable decisions, so data veracity is essential.

2. Value: Collecting, managing, and analyzing data can be expensive, but the knowledge therein can be quite valuable. Big data simultaneously increases the potential value and complexity of extracting it. The trick is recognizing the value, teasing it from data, and monetizing it.

3. Visualization: Transforming numbers into interactive charts and graphs for a wide variety of technical and non-technical users is critical for navigating, understanding, and using the data. Although this is true for all datasets, it's particularly relevant for big data due to the breadth of users and applications.

Applying Big Data to the Printing Industry

Even though the printing industry doesn't generate petabytes of data daily, big data could still play an important role. PSPs need to understand their business operations to maximize operational efficiency, reduce risk, and make better business decisions. Data sources and quantities continue to increase as PSPs digitally accept, process, and manage growing numbers of jobs. In the spirit of big data alliteration, I will categorize the three primary uses of big data by printers as presses, processes, and people.

Presses

Predictive analysis to identify preventative maintenance for presses is incredibly valuable. Looking at errors and error rates, such as misfeeds or recalibration rates, can identify potential problems before they arise, thus creating opportunities for preventative maintenance to avoid shutdowns. It can also help business owners determine when their machines are reaching the end of life and when their presses are over- or underutilized. Vendors can perform some of this predictive analysis using traditional data processing methods for individual shops. However, using anonymized big data across all presses in the field dramatically improves the ability to identify potential issues and best practices.

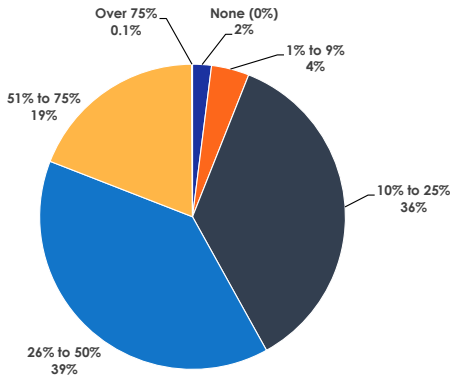
Processes

Perhaps someday, all cars will be self-driving, all offices will be paperless, and all printers will run as lights-out operations. The world isn't there yet, but the pace of business has increased. To remain competitive, we must understand where and how to streamline the processes that are not fully automated.

This problem falls squarely in the big data space. It requires a variety of data sources such as e-commerce, Management Information Systems (MIS), and Digital Front Ends (DFEs). The problem involves velocity to determine the immediate bottlenecks that are idling presses or loading docks so that production managers can reprioritize, reschedule, or redirect work immediately. Of course, managers must also perform deeper analysis to identify, understand, and address recurring process inefficiencies.

Figure 1: Percentage of Fully Automated Jobs

What percentage of your jobs are fully automated at this time?



N = 238 Respondents
Source: *Western Europe Production Software Outlook*; Keypoint Intelligence 2022

People

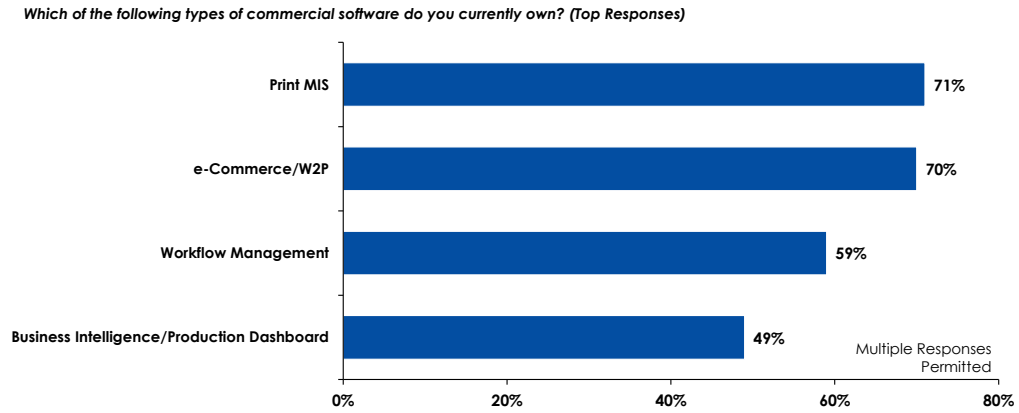
No one wants software constantly looking over their shoulders, but performance management is necessary, particularly in an operational setting. This is exacerbated in today's printing environment, where there is a mix of senior staff who might be more comfortable with traditional processes alongside new hires who might not have sufficient training. Simple analytics might show that one location or shift is less productive, but determining why can be challenging. Is the problem due to staff issues, job characteristics, or competing tasks? Supervisors can spend hours wading through job logs or physically managing staff to determine whether there are staffing issues and how to address them if they exist. Big data analysis can help uncover those answers.

Where Are We Today?

Most PSPs are already capturing data from a variety of sources, such as:

- E-commerce Storefronts
- Digital Front End (DFE) Job logs
- Production Management Solutions
- Management Information Systems (MIS)
- Enterprise Resource Planning (ERP) software
- Customer Relationship Management (CRM) systems
- Inventory/Fulfillment Management solutions
- Campaign Management Solutions
- Homegrown Spreadsheets and Relational Databases

Figure 2: Commercial Software Ownership Among PSPs



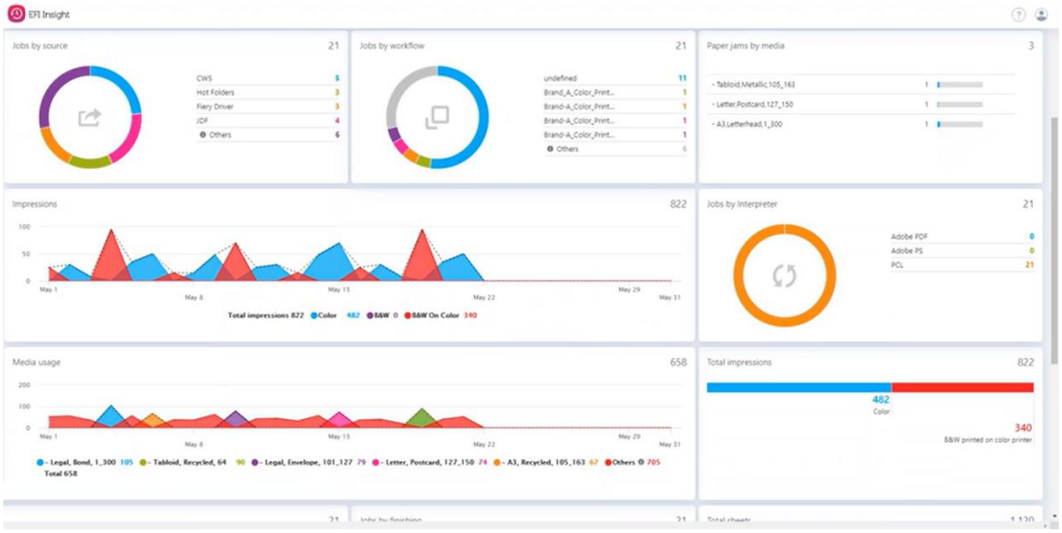
N = 238 Respondents
Source: *Western Europe Production Software Outlook*; Keypoint Intelligence 2022

The data from these various systems is stored in silos that rarely interact and are challenging to analyze together. However, production automation vendors are beginning to offer business intelligence and production dashboard analytics solutions. A few examples are described below.

EFI

EFI provides two print production analytics products called EFI IQ and EFI Insight. EFI IQ is included with Fiery servers and provides near, real-time visualizations of printer utilization and other printer statistics. The EFI Insight solution takes this quite a bit further by providing tools for analyzing this data over time to determine trends and comparisons between sites, groups of presses, and even shifts.

Figure 3: EFI IQ and Insight

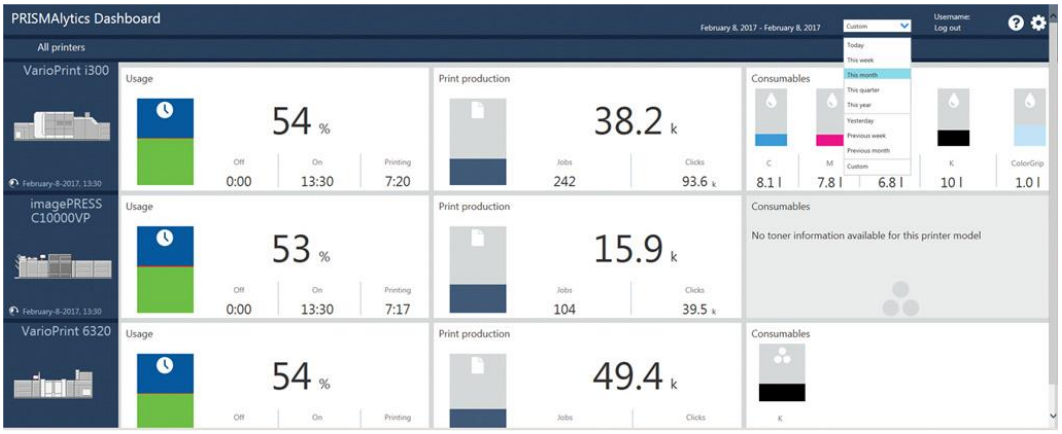


Source: EFI

Canon

Canon PRISMAlytics is a production print analytics package that is primarily designed to work with Canon's PRISMAsync DFE. Being linked to PRISMAsync, Canon PRISMAlytics can potentially provide more detailed information about Canon presses than other DFEs can access. It provides up-to-the-minute, interactive graphs of press volumes and activity and long-term trend analysis tools. The HTML5 interface is customizable, and many charts are interactive to bring up additional levels of detail.

Figure 4: Canon PRISMAlytics



Source: Canon

Ricoh

RICOH Supervisor is Ricoh's business dashboard/production analytics offering. Its customizable interface provides current shop activities and volumes as well as longer-term

volumes and trends for presses, locations, customers, etc. RICOH Supervisor is designed from the ground up to be vendor-agnostic.

Figure 5: Ricoh Supervisor

Source: Ricoh

It is also worth mentioning that Ricoh has a big data solution called RICOH Analytics for Print, which is designed for analyzing departmental printers across an enterprise. This solution incorporates cloud-based software, big data analysis, and professional services. Although it is outside the scope of this paper, it's still noteworthy.

All three of these solutions have similar capabilities:

- They are cloud-based with customizable interfaces
- They provide near-real-time volume, status, and performance data
- They collect and display information from one or many presses and locations
- They provide visualization tools for historical performance and trends
- They can provide several levels of detail, down to the actual job file data

They also have similar shortcomings:

- They are primarily data collection and visualization tools with limited analysis capabilities. These solutions can be quite helpful in highlighting anomalies but are somewhat frustrating in understanding the root causes (e.g., whether poor shift or press performance is due to hardware problems, operator issues, job characteristics, or other influences).
- They are focused solely on press operation and don't include other printshop operations such as onboarding, prepress, finishing, packing, or shipping. That's not an issue with these solutions; it's simply a limitation because all their data comes from DFE job logs. Data from functions outside DFEs is found in systems like e-commerce, MIS, workflow management, or fulfillment solutions. However, that

focus means these production dashboards are tightly focused on press operation, not the overall production workflow.

- MIS products can contain broader production data if operators' tracking capabilities are enabled and utilized by various stations. However, many PSPs primarily implement MIS for estimating and pricing and don't consistently track work through shops. Even if they do, detailed press data like that used by the dashboard solutions are rarely captured in the MIS.

Solutions, Vendors, and Tools

Global Big Data Print Solutions

Vendors will develop these solutions for their own use to present anonymized big data for use by other systems (e.g., determining the performance data for all installations of a specific press model by use or demographic). Vendors can use data like this to generate global performance benchmarks that local systems can use for comparison. These systems will help press vendors improve predictive service analyses and customer satisfaction. The data in these systems must be anonymized to remove any identifiable PSP, end customer, or job-related information.

Local Big Data Print Solutions

These solutions are limited to localized environments and provide detailed information about jobs, processes, and customers to drive PSP business decisions. They may tie into global big data solutions to benchmark how a specific provider, process, or press compares to industry norms. These can identify opportunities for business owners to streamline their operations and must access data from various sources, not just job logs.

Such solutions will likely be architected around the concept of data lakes rather than data warehouses to be efficient. Data warehouses are structured databases with rigid, predefined schemas that must be followed when storing data. Data lakes capture all sorts of data while the schemas, quantities, scope, and other attributes are defined when the data is read. Vendors might also use partially anonymized local big data solutions where client and job names are obscured for predictive maintenance analysis.

Press & Software Vendors

The ability to collect and analyze performance data across the entire marketplace and use that along with specific PSP machine data has tremendous service and sales implications. When developing these types of solutions, consider the following:

- Big data application development requires a different skill set than traditional database programming.
- Solutions must be extended beyond visualization to deeper analysis.

- It's important to query existing dashboard and analytics users about how they use current systems and what they try to learn. Direct your engineers to use this to design more efficient tools.
- Value impacts (preferably monetary) should be provided when identifying issues or suggesting improvements to help users prioritize.
- Data should be gathered from multiple sources such as DFEs, MIS, and other systems to provide accurate business dashboards across the entire production and delivery process rather than just press dashboards.

Print Service Providers

These sorts of tools have the potential to help you streamline your operations and increase profitability. These benefits will only grow as more advanced solutions come to market, but solutions are available now that can help visualize and analyze data sets. It's important to begin using the business analytics tools that are available today. MIS or workflow management solutions should be used to manage jobs through your shop rather than paper job tickets, job jackets, and job boards. It's time to start collecting the data!

opinion

The Bottom Line

The industry needs some big data solutions that look at the broader picture within printshops and across the marketplace. These solutions need to go beyond data capture and visualization. To be effective, they need to provide analysis and recommendations with value. This means that these future solutions need to go beyond visualization to assist people in locating issues, identifying root causes, and presenting potential monetary impact. For example, perhaps a PSP finds a press performing 30% worse than other devices. Does this press' performance even matter if there is a job backlog in prepress or onboarding? What are the financial implications of this information? There is not one big data solution in this space. Some solutions will focus on local PSP environments, while others will analyze the broader marketplace. Even so, both will be able to leverage each other's knowledge.

author

Greg Cholmondeley

Principal Analyst

+1 561.866.1384

Greg Cholmondeley is the Principal Analyst for Keypoint Intelligence's Production Workflow Consulting Service, which helps vendors define their future through consulting, market analysis, research, and forecasting. He also works directly with print service providers to improve their operations through workflow audits based on workflow journey mapping and the five stages of smart print manufacturing.

Comments or Questions?

Download our mobile app to access our complete service repository through your mobile devices.