O'REILLY®

Advancing Procurement Analytics

Capturing the Long Tail with Simplified Data Preparation

Federico Castanedo

Strata+ Hadoop WORLD

Make Data Work

strataconf.com

Presented by O'Reilly and Cloudera, Strata + Hadoop World helps you put big data, cutting-edge data science, and new business fundamentals to work.

- Learn new business applications of data technologies
- Develop new skills through trainings and in-depth tutorials
- Connect with an international community of thousands who work with data

Job # D2044

Advancing Procurement Analytics

Capturing the Long Tail with Simplified Data Preparation

Federico Castanedo

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Advancing Procurement Analytics
by Federico Castanedo

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Shannon Cutt
Interior Designer: David Futato
Cover Designer: Randy Comer
Illustrator: Rebecca Demarest

June 2016: First Edition

Revision History for the First Edition
2016-06-28: First Release

978-1-491-95611-3

[LSI]

Table of Contents

Advancing Procurement Analytics

Introduction

The explosive growth of data is enabling managers to make decisions that can give companies a competitive advantage. At the same time, making sense of this influx depends on the ability to analyze data at a speed, volume, and complexity that is too vast for humans, or for previous technical solutions. Organizations are challenged with not only surpassing their competitors, but making decisions to optimize their own business activities and workflows. Yielding insights from data has the potential to transform companies' internal processes and reduce costs.

An important area where this transformation has a huge business impact is the optimization of *procurement processes*. During the procurement process, some companies may spend more than *two thirds of revenue* buying goods and services, which means that even a modest reduction in purchasing costs can have a significant effect on profit. From this perspective, procurement—*out of all business activities*—is the key element in achieving cost reduction.

In a nutshell, procurement is about planning the buying process in a proactive and strategic approach. The process includes preparation and processing of a company's demand, as well as the end receipt and approval of payments. The process can begin by issuing a purchase order, and end when the order is shipped; or, it can cover a broader scope, which includes demand planning and inventory optimization. Demand planning and inventory optimization tasks are mostly data driven, and their outcomes depend on the *quality* of the input data and on the accuracy of the predictive algorithms.

The importance of procurement teams is clearly evident. In 2015, a Global Chief Procurement Officer Survey by Capgemini Consulting revealed that 72% of procurement groups reported to a C-level executive (in 2012/2013 it was a 59%), and more than 16% reported directly to the CEO. A study from IBM shows that companies with high-performing procurement teams report profit margins of 7.12%, as compared to 5.83% from companies with low-performing procurement teams. In addition, companies with top-performing procurement teams report profit margins 15% higher than the average performing company, and 22% higher than low performers.

Locate, Categorize, and Maintain Data

To generate savings faster than their competitors, procurement teams should have an appropriate way to locate, manage, and maintain data; the challenge, however, is that data is not always easy to collect because it is usually spread throughout the organization.

Traditionally, procurement organizations have the goal of maximizing cost savings, and to achieve it they usually focus on the spend of the top suppliers. This approach is based on the Pareto 80/20 principle: approximately 80% of the spend will be covered by 20% of the suppliers; on the other hand, the remaining 20% of the spend is covered by the other 80% of suppliers. Nevertheless, in some cases the long tail can be 50% of the total spend by the organization. It is common to focus on the top suppliers rather than analyze the complete long tail, because sourcing managers do not have enough time. But if the time spent in the process of analyzing data can be reduced, it will be possible to analyze the *complete* long tail and take advantage of the complete picture (Figure 1-1).

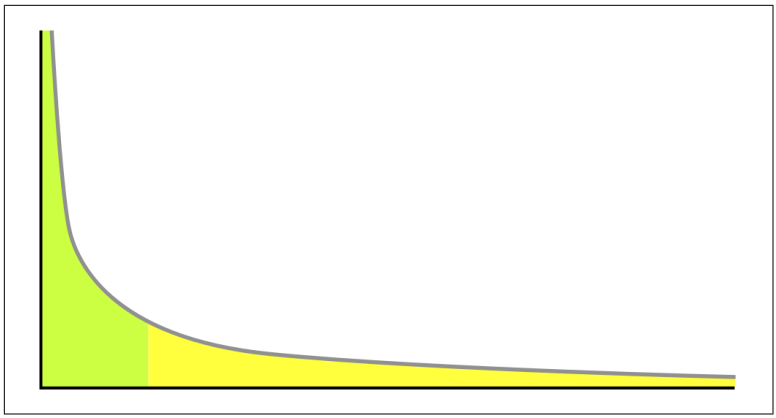

Figure 1-1. Supplier/buyer's spend usually follows a Zipf distribution. The long tail in yellow may have an amount higher than the green one but is split over a high number of suppliers.

Overcoming Unexpected Events

Procurement or sourcing managers need to purchase the right quantity of products at an advantageous price and at the right time. Therefore, it is important to understand how delays, disruptions, and other unexpected events affect the overall operations and the sourcing costs. That means managers need to be fully aware of the potential impact of geopolitical and other events in the demand of the products they need to acquire.

To overcome unexpected events, managers need instant access to a supplier database to identify new suppliers if necessary. A key consideration is to have immediate access to the profile of trusted supplier data, enabling a buyer to start commercial transactions with new suppliers. As an example, blur cloud software provides a web application to transparently and simply manage, source, and deliver services. It allows the user to create project briefings and use the blur marketplace with more than 65,000 service providers. Other startups, like Tradeshift, focus on simplifying the invoicing operation by providing a supplier platform for invoices and payments, using connections between companies to verify the transactions in a manner similar to social networks. Other companies focus on streamlining the entire procurement process using cloud-based solutions, like Ariba and Taulia.

Leading procurement organizations are also augmenting their information with trusted third-party sources to respond efficiently to unexpected events. As an example, Tamr's platform provides integration with Reuters data, allowing the analysis of the supplier market and the ability to track significant news (e.g., bankruptcies).

Procurement in the Public Sector

Procurement is also an important topic in the public sector, where there are potential benefits for the government. In most countries, it is also mandatory to publish the public contract notice to ensure enough transparency. As an example, the website OpenProcure lists US public agencies and their respective procurement thresholds; these thresholds identify the dollar amount under which a government agency can purchase a product without the requirement of doing a competitive bid.

Data integration of public contracts is a related topic in the European Union. Public contracts must be available by law in the EU, but data is not easy to obtain, and published data commonly appear in different formats and languages. Lod2 is a large-scale research project funded by the European Commission with the goal of advancing the representation of public contract data to enable electronic data integration. They propose that public contracts can be represented using *linked data*—allowing semantic queries and links to external information.

Current Solutions

In today's big data era, procurement teams want to be more data driven, and data sources cannot be managed as a group of individual silos. As procurement teams begin to collect and maintain higher-quality data, advanced analytics techniques will be utilized to drive decision-making strategies and identify opportunities.

Most procurement organizations have some data infrastructure in place. Typical infrastructure components are *Enterprise Resource Planning* (ERP) systems, which primarily manage direct spend with suppliers, and *Source-to-Pay* (S2P) systems that manage *indirect* spend with suppliers. Some basic analytics, focused primarily around spend, are usually performed with this software to answer business questions.

Spend Analysis

Spend analysis is the process of collecting, cleaning, classifying, and analyzing procurement data with the purpose of decreasing costs, improving efficiency, and monitoring compliance. There are many benefits of spend analysis and management, such as reductions in materials and services costs, inventory costs, decreased sourcing cycle times, and improved contract compliance. The cost, lack of knowledge, or availability of scalable spend analysis tools are common roadblocks.

Data-Driven Action

The original approach to analyzing spend is to build "spend cubes" along three dimensions—(1) suppliers, (2) corporate business units, and (3) category of item—where the contents of the cube are the price and volume of items purchased. Using procurement analytics to determine things such as how much is spent by supplier, category, etc., can lead to the following data-driven actions:

- **Aggregation:** It is possible to reduce the supplier base and increase the cost savings by the aggregation of multiple suppliers for a single product. This provides direct savings based on the difference among current prices and negotiated contract pricing.
- **Compliance:** Discover contracts that should be carried out following specific terms, but for whatever reason were not accomplished; this includes monitoring the terms and conditions of the contractual agreement and tracking rebates and payment terms.
- **Untouched spend:** It may be the case that high costs in some categories go unnoticed by the procurement team. This may happen because managers do not have enough time to analyze all of the categories and existing tools are not quick enough.
- **Price arbitrage:** This happens when multiple prices are charged for the same unit even from the same supplier. Price arbitrage requires having the right information at the right time and enables you to estimate costs before quotes are received.

- **Spend recovery:** This allows you to detect duplicated invoices for payments, whether done intentionally, as in the case of fraud (example from Boeing), or not.

Managing Costs at a Sub-Commodity Level

To understand and identify the true drivers of cost in a big organization, it is necessary to manage costs at sub-commodity level, using detailed taxonomies. This process involves diagnosing price differences of similar components by integrating several data sources, and it allows businesses to make decisions at the sub-commodity level.

To identify key suppliers to partner with, it is necessary to understand sales, trends, and growing/declining product lines; it's also necessary to monitor and analyze market developments. A critical factor for success is not only having access to *all of the data* from the different subsystems, but also having high-quality, accurate data. Moreover, to be able to react on time, the procurement analytics actions should be carried out frequently—not only once or twice a year. Finally, the analytics results must be easy to use in order to make the right decisions.

As an organization becomes more mature and grows, problems with procurement analytics arise, limiting their ability to quickly and effectively answer business questions and generate adequate data-driven actions. These problems primarily revolve around data preparation and can be classified as:

- Lack of quality in data preparation, due to data variety.
- Speed of data preparation.
- Lack of scalability in data preparation.

We will focus on these problems, and how they can be addressed, in the sections that follow.

Dealing with Data Variety

Sourcing managers usually have both *quantitative* and *qualitative* data, with different formats. Before doing any type of analysis, this data must be prepared and integrated, or *curated*, to represent accurate information.

As companies struggle with the *amount* and *variety* of data stored, they find it difficult to centralize and integrate it in one place. This situation especially arises in large corporations, which often have systems from different vendors and data stored in different formats (resulting in *data silos*). Large and mid-size organizations may have five or more sources of spend data. Furthermore, legacy vendors do not have sophisticated automation techniques for data preparation and require human involvement.

Broadly speaking, there are two solutions for the data variety problem:

1. Embark upon a complete transformation of all the software platforms and databases, and generate the data into a common format/schema.
2. Use an integration and data unification platform.

In procurement, data *variety* often appears when you have business units in different countries. For example, it may be the case that a business unit with offices in both Spain and France has different ERP systems, where the same item may be stored using different IDs. Most of the time, this occurs because the supplier provides different IDs for the same item, and possibly different pricing as well. So the internal ERP system records the ID provided by the local supplier and does not have visibility of other countries' data. Another example is within a Supplier-to-Procurement system (S2P), where there may be many entries related to the same supplier. For instance "General Electric" may be also be entered as "GE," "Gen," "Gen Electric," etc. All of these different entries for the same entity lead to confusion and wrong analytics results. It is common to have a lot of records that need to be assigned/classified into a material group or commodity code. This classification of things into broader categories—for example, in building a catalog—is something that can be automated very efficiently using *machine learning algorithms*.

Universal Business Language

Undertaking data integration to overcome data variety is a well-known issue in computer science. Several languages, such as XML, have been proposed to develop *middleware* layers and enable data integration. To solve the integration problem in B2B, the OASIS Universal Business Language (UBL) was developed. It defines a

generic XML interchange format for business documents, which can be used to meet procurement requirements. One of the drawbacks of XML is the required data overhead, due to the fact that its foundation is built on using tag pairs to represent elements. Currently, UBL is being replaced by JSON encoding, which provides a lightweight approach to integrating data.

For more information about the technical issues of data preparation, we refer the reader to the free O'Reilly report, *Data Preparation in the Big Data Era*.

Speed and Lack of Scalability in Data Preparation

While it's clear that it's very important for organizations to operate quickly, analyzing massive amounts of data quickly is a major challenge. Existing solutions often require manual approaches to integrate and clean data, are often cost and time prohibitive, and prevent organizations from scaling to more sources. Given this situation, procurement analytics are generally focused on only a *fraction* of the available data. Cleaning and joining data using conventional methods, even before using any analytics tools, can cause reporting to take weeks to months to generate.

Sourcing managers need to make decisions based on spend analysis. One of the objectives of spend analysis is to support strategic sourcing and cost reduction initiatives. It is necessary to have a general view of the company's spend in order to understand overlaps in supply chain and purchases. This means that it's critical to boil the data down into something that can be acted upon in a reasonable timeframe, to either help companies generate more revenue, serve customers better, or operate more efficiently.

Novel Approaches to Procurement Analytics

Most organizations rely on ERP data and Excel to run the majority of their analysis for procurement. This often involves multiple people working on the same dataset—creating massive inefficiencies. In addition, scaling the operation under these conditions creates an exponential cost curve. Even procurement legacy vendors do not have sophisticated automation techniques for data preparation and integration, so manual effort is still required. These approaches do

not scale well because they need human intervention to solve data integration issues.

A higher level of automation is possible with machine learning algorithms that automatically interact with the user to solve the integration problems jointly. This new approach should provide the benefits of increased speed and scalability of the complete data preparation operation, including cleansing, integration, and classification of datasets. This leads to faster answers, fewer "fire drills," greater visibility into parts or suppliers, and enhanced trust in the analytics process.

One example is the Tamr platform, which is a tool designed to simplify the data preparation and unification process. The platform builds a global view and allows the user to generate reports and data analysis. It provides a probabilistic, bottom-up approach to the complete data preparation operation, leveraging automation and human input in the process of validating data. The Tamr platform is also capable of connecting with different systems and data sources (even third-party data) and automatically builds a taxonomy. It can also be used to *migrate* data from legacy systems to ERP and can integrate and unify data to generate a clean dataset for migration. Several examples of sourcing analytics dashboards generated from the Tamr data integration platform can be found on this site.

The machine learning capabilities of Tamr reduce time for data integration, allowing the organization to scale. The platform also has the ability to accept expert feedback—helping the user handle exceptions and conflicts in the system. Although there are other software tools that automate the procurement process (like BellWether and BravoSolution), they are not prepared to work with existing legacy systems.

By automating and reducing the amount of time required for generating reports, managers can spend their time in negotiations with suppliers, rather than working on reports, allowing them to analyze more data quickly and uncover more opportunities. The idea is to allow deeper analysis with fewer resources, or at least without adding more.

Opportunities in procurement are not always easy to detect and may be subtle. As an example, a spend analysis report from Concur, the automatic travel expense management software company, highlights masquerade purchases, duplicates, and out-of-pocket expenses as

the greatest areas of concern. Through spend analytics, Concur was able to determine an interesting figure: by crunching 10M transactions, they detected that employees who purchased in-room movies tended to spend less overall on their trips.

The Next Step Forward

Novel and intelligent software solutions are enabling procurement organizations to make faster and more effective decisions. By using the correct tools, extracting core ERP data and combining it can take minutes, when it previously took days. New procurement solutions will enable automatic aggregation and analysis of data from diverse sources and will provide an efficient view of the dispersed information. Ideally, these new tools will provide automatic notifications of risk, saving opportunities, and suggest improvements in supplier relationships—but we are not there (yet).

Game Theory

Procurement has also been a research topic in academia from the game theory and auctions perspective. *Game theory* applies mathematical models to the process of decision making, in order to predict the outcome of the interaction. The application of game theory to the procurement process can be used to understand how and when the buyer can increase the pay-off in their favor (by reducing the price). In their paper "Truthful Multi-unit Procurements with Budgets," Hau Chan and Jing Chen presented research for the *bounded knapsack problem*—a special class of procurement games where each seller supplies multiple units with a cost per unit known only to him. The buyer can purchase any combination of units from each seller, under a specific budget. It has been shown that for multi-unit settings with budget considerations, no mechanism can do better than an *ln n-approximation*, where n is the total number of units of all items available.

Inventory Optimization

Inventory optimization or management is another well-known research topic. In their paper "Optimal Dynamic Procurement Policies for a Storable Commodity with Lévy Prices and Convex Holding Costs," Chiarolla et al. discuss inventory management policies in

the presence of price and demand uncertainty. They focus on the inventory of a commodity traded in the market, whose supply purchase is affected by price and demand uncertainty.

More related research can be found in The Journal of Purchasing & Supply Management.

Machine Learning in the Future of Procurement

In the future, we will see more applications and use cases of machine learning to improve and optimize procurement practices in order to reduce costs and increase margins. One area likely to show advances is unsupervised machine learning or clustering to detect similar contracts. In this application, insights have the potential to identify the most suitable contracts, help prepare contracts by detecting those that are similar, and identify opportunities from demand aggregation. Predictive models can be also used to infer the number of bidders for a public contract. In this situation, the bigger the number of bidders that will be estimated for a contract, the better. It can also help to detect suppliers that can offer the same service or products (for an interesting example, check out this kaggle competition sponsored by Caterpillar, where the goal was to predict the price a supplier will quote for a specific tube assembly). For even more information about procurement, we refer the reader to the Chartered Institute of Procurement & Supply (CIPS).

About the Author

Federico Castanedo is the Chief Data Scientist at WiseAthena.com, where he analyzes massive amounts of data using machine learning techniques. For more than a decade, he has been involved in projects related to data analysis in academia and industry. He has published several scientific papers about data fusion techniques, visual sensor networks, and machine learning. He holds a Ph.D. on Artificial Intelligence from the University Carlos III of Madrid and has also been a visiting researcher at Stanford University.