research

CHECKLIST REPORT

2017

Data Monetization:

7 Steps to Building Consumable Data Solutions

By David Loshin

Sponsored by

Google Cloud

tdwi

Transforming Data With Intelligence™

DECEMBER 2017

TDWI CHECKLIST REPORT

Data Monetization:

7 Steps to Building Consumable Data Solutions

By David Loshin

TABLE OF CONTENTS

tdwi
Transforming Data With Intelligence™

555 S. Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org

FOREWORD

Lower barriers to entry to creating systems for acquiring, ingesting, processing, and analyzing massive data volumes have allowed a multitude of organizations to embrace business intelligence and analytics. The spirit of data democratization inspires organizational leaders to empower a wider range of team members by providing them with the information they need to make the right decisions. However, even with the simplicity of modern end-user reporting and analysis tools, there are still opportunities to build data solutions that are more easily consumed by the different target audiences.

Data products are not a new idea; data aggregators have been producing purchasable data sets for decades. However, as organizations have become motivated to be "data driven," the concept of a "data product" has rapidly morphed into different shapes, including packaged data sets, lightweight API-based services, directly connected end-user visualizations, and full-blown access to hosted reporting and analytics dashboards.

Innovative methods of packaging and selling information are inspiring organizations to exploit their data and create new sources of income. This is the essence of data monetization—the ability to fuse data from a variety of sources, configure a data product that addresses business concerns, and deliver it as a value-added (and revenue-generating) data solution.

A "data solution production factory" can address the information product expectations of a continuum of downstream consumers, including internal customers who consume solutions, trusted partners with whom your organization shares information, existing clients hungry for additional insights from their own data sets, or a much broader audience of external customers eager to pay for your information. This Checklist Report looks at some of the steps in creating an environment for producing consumable data solutions.

However, these solutions must go beyond the conventional data extract. They must be able to ingest and process high-speed data streams, fuse data from a variety of internal and third-party sources, scale with the volume and velocity of incoming data, and provide multiple delivery methods (such as portals, dashboards, or direct access) to different consumers. Assembling a plan for streamlining the delivery of data solutions can be a profitable endeavor that is increasingly simplified when partnering with cloud-based business intelligence/analytics providers.

NUMBER ONE

ASSEMBLE A DATA SOLUTION PLAN

The data product of the past is woefully inadequate to meet the diverse needs of today's potential consumers. A modern data solution must not only package the information to be consumed, it must also encompass the means by which the information is delivered, the methods of data presentation, and the ability to further integrate within other downstream operational and analytical processes.

Producing value-added data solutions requires forethought and planning. The first step is to develop a template plan for configuring, designing, producing, and delivering a data solution. There are multiple goals: simplifying the engagement process, streamlining the methods of understanding what information is needed to support and improve internal data consumer business processes, clarifying the kinds of data solutions that can be marketed to trusted partners, and determining what data solutions can be made available to broader audiences.

The steps include (but are not limited to):

- **Identify opportunities for data monetization.** Recognize the potential value of fusing data from a variety of sources and consider the practicality of configuring and delivering different types of data solutions.
- **Identify your data consumers.** Determine the communities of customers that can benefit from a data solution.
- **Engage your prospects.** Envision how the data solution meets consumer needs and consider ways of engaging and partnering with the potential customers.
- **Solicit requirements.** Assess the data consumers' needs and solicit the information necessary for more precise planning.
- **Plan your product.** Evaluate a variety of solution packaging, presentation, and delivery alternatives.
- **Identify data sources.** Determine what source data is required to produce the data solution.
- **Data production flow.** Design the processes for data ingestion, preparation, transformation, and integration into the analytics pipeline that produces the data solution.
- **Data solution delivery.** Provide wireframe mockups of the methods for delivering the data solution.
- **Production.** Develop agile processes for producing and delivering the data solution.

Recognize that not all organizations are identical. One size does not necessarily fit all. Customize your plan based on your organization's best practices, the types of data consumers, and the types of data solutions to be produced.

NUMBER TWO

DESIGN TO INGEST AND MANAGE HIGH-VELOCITY DATA

We are already inundated by data, yet the pace of data creation continues to accelerate. Data volumes continue to explode as social networks grow, creating additional streaming sources of human-generated data. In addition, emerging Internet of Things (IoT) applications are expected to create streams of machine-generated data. Organizations that can handle massive volumes of high-velocity data are positioned to analyze these streams to create downstream data solutions. The critical point to recognize is that the value of the delivered information product is enhanced when it can be delivered to the right people in a timely manner.

Organizations will want to integrate data streaming from a wide array of sources in real time. Quickly delivering actionable knowledge increasingly implies that your data solution production system must be designed to ingest data at high speed and consequently process, persist, and analyze it in preparation for inclusion in an end-user data solution. At the same time, if possible you want to avoid having to continuously monitor and adjust the environment to be able to scale with growing data volumes and velocity.

When designing a framework for producing data solutions, it is important to ensure that the results of real-time analysis of streamed events are immediately visible to all consumers of produced data solutions, including internal analysts, operational systems, and external partners.

In addition to the usual functional requirements for ingesting, managing, and analyzing data, consider these system requirements to support high-volume and high-velocity streaming data:

- Data ingestion using different data processing patterns that balance batch and continuous computation, depending on the incoming streams
- Computational performance that limits end-to-end latency and does not introduce delays into the process flow
- Dynamic computational scaling in relation to the speed and volume of all the data streams
- Dynamic storage scaling that allows the system to expand and contract storage requirements as necessary
- Data mapping of incoming data to define schemas that ensure data consistency
- Direct integration with analytics models that can facilitate updates to the data solution

These concepts will help guide your design for ingesting and managing high-velocity data.

NUMBER THREE

CROWDSOURCE ANALYTICS USING INTERNAL, PUBLIC, AND COMMERCIALLY AVAILABLE DATA SETS

Using externally provided data to enhance your internal business intelligence and analytics is not a new idea. For many years, research companies and data aggregators have collected and integrated data from a variety of sources to create and sell packaged data products.

However, today the potential for producing usable data solutions by integrating external data is greater than ever before as a result of three key environmental factors. First, low-cost virtual and cloud computing systems simplify the creation of your own data production pipelines. Second, numerous publicly available data sets (such as government agency-published open data sets and licensable, commercially available data sets) can be fused with internal data to create enriched, value-added data products.

More important, though, is the third factor—newer operational models for ingesting externally produced data that, under the right circumstances, allow that externally produced data to be delivered directly into user projects. This is often done by connecting data stream pipelines directly into cloud-based analytics systems or data warehouses. For example, given the right cloud data warehouse provider, you can create a set of pipelines that connect to third-party data providers (such as social networking streams, point-of-sale data, or transactions for hosted ad networks) and have streaming transactions stored directly into a data warehouse.

One other facet of hosted cloud-based environments is the potential for cross-client data sharing. Client data environments are typically segregated within multitenant environments, but in certain scenarios, you might be able to leverage the aggregation of data from multiple client sources even if that source data is protected. For example, you might roll up product purchase statistics for all suppliers by product category without exposing how much of each supplier's brands were sold, or your host provider could facilitate negotiating data-use agreements among their clients to allow for additional shared data streams.

Combining internal data sets, external data products, and a variety of real-time data streams provides a fertile base for crowdsourcing

a range of end-user data solutions. The result is a richer palette to support the creation of data solutions.

NUMBER FOUR

INTEGRATE AUDITING TO COMPLY WITH DATA OBLIGATIONS

The open data movement has motivated transparency in many (mostly government) organizations to regularly produce data. In addition, a growing number of commercial companies expose selected data sets in different ways, such as via flat file downloads or streamed through Web services.

Consuming externally provided data does come with caveats, especially when the data is proprietary. Government agencies may request that data consumers abide by stated use policies or require signing data-use agreements that specify how the data may or may not be used. Private sector companies complying with regulations and industry guidelines for protecting sensitive data will likely insist that shared data be subject to compliance with their well-defined obligations.

Consider a company that shares customer data with a trusted partner. That trusted partner may be restricted from using a customer's data for sales purposes unless that partner already had a relationship with that customer. Other imposed obligations may include geographic data use restrictions, the need to monitor which individuals or systems touch personal information, ensuring encryption of sensitive data, maintaining an audit trail for record handling, and abiding by archival and disposition directives in a timely manner.

When data sharing is safeguarded by accompanying obligations, it is important to not just comply with those data obligations, but to have an auditable process to demonstrate that the system conforms with any data protection or oversight directives. Therefore, build the proper auditing, oversight, and compliance reporting into the environment used to produce and deliver data solutions to the customer community.

Specific methods include:

- **Encryption.** Incorporate methods for encrypting data at rest and in motion (to prevent exposure of sensitive data in the event of a security breach) and generate notifications when data is encrypted or decrypted.
- **Geographic storage.** Demonstrate that data is stored in specific locations to abide by data location restrictions.
- **Access control.** Exercise full control over specific roles that are allowed to access sensitive data and document when sensitive data is touched.
- **Business rules.** Integrate monitoring and reporting of compliance to defined business rules.

NUMBER FIVE

BUILD SCALABLE AND EXTENSIBLE SYSTEMS

Emerging data solutions and corresponding data products are increasingly dependent on processing and analyzing massive data volumes, and the underlying system architecture must be adequately provisioned to support the computational and storage needs. However, because conventional hardware systems are sized (and priced) according to required performance objectives, it may be difficult to anticipate computational and storage needs. This is especially true when the analytics that drives the creation of downstream data products may change as more data sources are added, data volumes expand, data velocity accelerates, or more consumers are served.

The conventional platform acquisition process is unlikely to meet your data monetization needs. Instead, use scalable hardware configurations that employ commodity components and provide a level of elasticity.

One approach is to design and build your own commodity-based scalable computing system on premises. When you design and build your own on-premises system, recognize that in addition to increased flexibility you face increased costs and levels of effort for acquisition, operations, and maintenance. You will also run the risk of obsolescence of the selected components, requiring ongoing hardware refreshes to maintain or improve performance.

A second approach is to engage a cloud-based platform-as-a-service (PaaS) provider to host and manage your warehouse and analytics environment. Aside from reduced costs and effort, this is an increasingly attractive option for a number of reasons, including:

- **Time-to-value.** There is no need for a lengthy acquisition process and configured virtual clusters can be up and running within minutes.
- **Seamless scalability.** Cloud-based systems can be designed to dynamically grow with computational, storage, and performance demands, simplifying scalability.
- **Functional extensibility.** Your organization can benefit from the cloud service provider's experience in developing and maintaining many kinds of systems with different capabilities. Cloud-based environments are designed to easily integrate with high-velocity data streams, data warehouses, and embedded analytics.

An extensible system whose resources can be seamlessly scaled up or down depending on the demand will reduce startup costs and support incremental growth in delivering data solutions.

NUMBER SIX

CONSIDER OPTIONS FOR DATA DELIVERY

There is a qualitative difference between what was once called a "data product" and what we today would characterize as a "data solution." The archaic method of delivery produced a data extract delivered to the client; once the files were transferred, it was the customer's job to load the files, verify the data, and manipulate the data sets into a usable format. Aside from imposing a significant burden on the data consumer, effectively delivering the data set by "throwing it over the wall" implied a failure of the data producer to be accountable for the quality and usability of the product.

Today's data solution provider is dedicated to streamlining the ability to deliver insights to its clients. Therefore, a modern approach to providing data solutions may incorporate a number of alternatives for delivery. Consumers may expect to tinker with data extracts, but a growing delivery trend is to provide moderated direct access to the data platform. This may range from providing access through a crafted portal with customized predefined reports to completely open access to the underlying data warehouse.

Consider the different options for data solution delivery. Evaluate your customers' technical savviness and determine whether they are ready for direct query access to the data warehouse environment. This provides flexibility in providing accessibility versus allowing each consumer to develop their own end-user applications.

At the same time, recognize that exposing the data to direct access may lead to increased computational demand on the data warehouse platform. To avoid this potential bottleneck, segregate the different types of processing from each other and separate data processing from data storage. Isolating ingestion and transformations prevents performance degradation for downstream data accessors.

In some cases, the same or similar reports and dashboards can meet the needs of different customers, especially those served in a cloud-based multitenant environment. Developing customized portals with integrated data access controls allows reuse of reporting and analysis templates in multitenant environments without exposing one customer's data to any of the others.

NUMBER SEVEN

EMBRACE A TRUSTED CLOUD PARTNER

Producing modern data solutions can be a profitable venture, but there can be a significant amount of effort and investment necessary to build a data warehouse environment that can ingest and process high-velocity data as well as scale its storage seamlessly and provide the computational performance to meet the needs of an expanding cadre of customers.

Although some organizations have the resources, skills, and discipline to design, develop, implement, and operate their own data warehouse, the lack of such resources should not be a barrier to data monetization. Developing a data warehouse in the cloud or partnering with a PaaS host provider may be a more attractive alternative than attempting to build a scalable, high-performance cluster on premises. Choose a cloud partner you trust and develop an analytics platform that can support your data solution production needs.

Some features to consider include (but are not limited to):

- **Product landscape.** As we have seen, the desired platform must encompass scalable computation, storage, networking, data warehousing, big data analytics, and real-time data flow for batch and stream processing.
- **Performance.** The production environment must be able to produce the desired data products in a timely manner to meet the service-level agreements set for your downstream clients.
- **Dynamic scalability.** Look for a provider that automatically scales both the compute and storage capabilities based on your application's processing demands.
- **Stream processing.** Connecting to real-time streams is an indispensible requirement for modern data solutions.
- **Value-added data feeds.** Cloud vendors that can bundle system functionality with value-added data feeds are preferable because they can supplement your data solution engine.
- **Security and protection.** Ensure that the host has the proper tools and functionality to institute auditable data protection.
- **Cost.** Different providers have different cost models; evaluate your usage patterns and work with vendors to understand how their pricing models translate into actual costs.
- **Reliability.** Seek out cloud providers with high degrees of system reliability and guaranteed uptime.
- **Integration with downstream tools.** Ensure compatibility with end-user reporting and data analysis/visualization tools.

AFTERWORD

Complex analytics is rapidly becoming mainstream and data monetization represents the next phase in data-driven business. As the barriers to big data analytics and high-performance computing are dropped, more organizations will be poised to enhance their data assets with a combination of batch data sets and a plethora of high-velocity streaming data sources.

Intelligent organizations recognize that the key to engineering a framework for mass-producing data solutions is to avoid the impulse to build your own environment and instead consider partnering with a PaaS data warehouse and analytics provider. With the benefits of cloud computing—lowered operations and management costs, a wide array of available functional technology components, integration with high-velocity data streams, and dynamic scaling—your organization can work with the service provider to develop a platform that can provide different methods for data solution delivery to internal and external data consumers.

ABOUT OUR SPONSOR

Google Cloud

cloud.google.com

Google Cloud Platform (GCP) makes business insights available on demand via a set of serverless data analytics services that surpass conventional limitations on scale, performance, and cost-efficiency.

You can leave the complexities of data analytics behind and

- Use Google BigQuery, a cloud-native serverless data warehouse that executes queries in seconds instead of minutes, at any scale, for accelerated time to insight
- Ingest and analyze up to millions of events per second in real time with Cloud Pub/Sub and Cloud Dataflow
- Get value faster from data processing on Apache Spark and Apache Hadoop with Cloud Dataproc
- Visualize and explore data, publish dashboard and reports to share insights using Google Data Studio and existing third-party BI tools
- Bring predictive analytics into your applications by adopting machine learning at your own pace using Cloud Machine Learning Engine or pre-trained machine learning APIs

Please visit https://cloud.google.com/solutions/big-data/ for more information.

ABOUT THE AUTHOR

David Loshin, president of Knowledge Integrity, Inc., (www.knowledge-integrity.com), is a recognized thought leader, TDWI instructor, and expert consultant in the areas of data management and business intelligence. David is a prolific author regarding business intelligence best practices, as the author of numerous books and papers on data management, including *Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph* and *The Practitioner's Guide to Data Quality Improvement*, with additional content provided at www.dataqualitybook.com. David is a frequent invited speaker at conferences, web seminars, and sponsored web sites and channels including TechTarget, The Bloor Group, and as practice director for data platforms at Eckerson Group. His best-selling book *Master Data Management* has been endorsed by many data management industry leaders.

David can be reached at loshin@knowledge-integrity.com.

ABOUT TDWI RESEARCH

TDWI Research provides research and advice for BI professionals worldwide. TDWI Research focuses exclusively on analytics and data management issues and teams up with industry practitioners to deliver both broad and deep understanding of the business and technical issues surrounding the deployment of business intelligence and data management solutions. TDWI Research offers reports, commentary, and inquiry services via a worldwide membership program and provides custom research, benchmarking, and strategic planning services to user and vendor organizations.

ABOUT TDWI CHECKLIST REPORTS

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.