



CONSULTANCY

---

# Strengthening Self-Service Analytics with Data Preparation and Data Virtualization

A Technical Whitepaper

---

Rick F. van der Lans  
Independent Business Intelligence Analyst  
R20/Consultancy

September 2015

Sponsored by



Copyright © 2015 R20/Consultancy. All rights reserved. Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. or there countries. To view a list of Cisco trademarks, go to this URL: [www.cisco.com/go/trademarks](http://www.cisco.com/go/trademarks). Trademarks of companies referenced in this document are the sole property of their respective owners.

## Table of Contents

---

1	Introduction	1
2	Phase 1 of Self-Service Analytics: IT-Driven Data Integration	2
3	Phase 2 of Self-Service Analytics: Self-Service Data Integration	4
4	Phase 3 of Self-Service Analytics: Self-Service Data Preparation	5
5	Phase 4 of Self-Service Analytics: Cooperative Data Integration	8
6	Scenario 1: Data Virtualization as Data Source for Data Preparation	10
7	Scenario 2: Using Data Virtualization to Make Answersets Developed with Data Preparation Available for Everyone	13
8	Scenario 3: Using Data Virtualization to Operationalize User-Defined Data Sets	16
9	Scenario 4: Data Preparation for IT Specialists to Develop CIS Views	18
	About the Author Rick F. van der Lans	20
	About Cisco Systems, Inc.	20

## 1 Introduction

---

*Self-service analytics* implies that users design and develop their own reports and do their own data analysis with minimal support by IT. Most recently, due to the availability of tools, such as those from Qlik, Spotfire, and Tableau, self-service analytics has become immensely popular. Besides powerful analytical and visualization capabilities, they all support functionality for accessing and integrating data sources. With respect to this aspect of data integration four phases can be identified in the relatively short history of self-service analytics. This whitepaper describes these four phases in detail and shows how the tools *Cisco Data Preparation (CDP)* and *Cisco Information Server (CIS)* for data virtualization can strengthen and enrich the self-service data integration capabilities of tools for reporting and analytics.

**Phase 1: IT-Driven Integration** – In the first phase of self-service analytics, users were given the *freedom to analyze*. They were totally free to query, report, and analyze the data stored in data marts and data warehouses. However, with respect to accessing and integrating data the users were not free. Most tools support features for integrating data sources, but because of the technical complexity this aspect was organized and managed by IT.

**Phase 2: Self-Service Integration** – In phase two, the users' freedom to analyze was extended with the *freedom to integrate* data sources. With the new built-in *data blending* features users themselves could integrate data from data marts, data warehouses, personal spreadsheets, data sets with social media data, customer data from marketing companies, and so on, without IT involvement. No restrictions applied with respect to integrating data sources for reporting and analytical purposes. However, the users had to understand the data models and data structures of the data sources to implement the correct integration.

**Phase 3: Self-Service Data Preparation** – As indicated, while tools with data blending features make it easy to integrate data sources, they still demand that users fully understand the data structures. In the third phase, *self-service data preparation tools*, such as CDP, were introduced. These tools *guide* users in understanding the data and the data structures. They don't have to study what the best way is to integrate two data sources or find out which data values are potentially incorrect. All this is done automatically by CDP. This guided form of integration makes data available for reporting and analysis even sooner.

**Phase 4: Cooperative Integration** – Currently, the market has moved on to phase four. Self-service analytics hasn't replaced all the other forms of reporting and analytics, such as standard reporting, embedded analytics, ad-hoc reporting, and mobile BI, but complements them. All these forms can be divided in two categories: *IT-driven BI* and *Business-driven BI*. Self-service analytics belongs to the second category. In phase four, self-service analytics *cooperates* with IT-driven BI. In other words, self-service analytics becomes a fully integrated part of the larger BI environment. Also, with this new form of self-service analytics the gap is bridged between the two BI forms. Reports initially developed by the business may have to be "moved" to the IT-driven BI environment afterwards. This is referred to as *operationalization* of a self-service report. Another way to bridge the gap is when reporting specifications developed by IT are shared by users of self-service tools. For example, the integration of data sources may be so complex that the specifications are developed by IT and handed over to the business for more interactive forms of reporting. In this situation, IT *enables* business-driven BI.

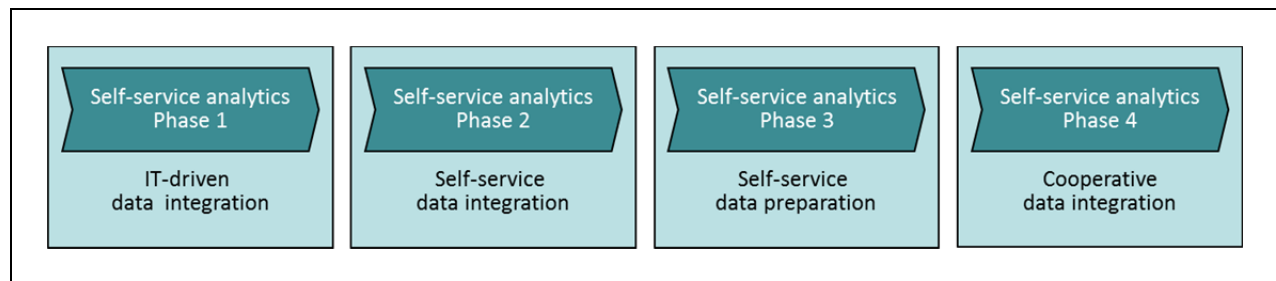
**The Need for Data Preparation and Data Virtualization** – To support both BI forms and to bridge the gap, two technologies are required: *data preparation* and *data virtualization*. In fact, data preparation and data virtualization strengthen self-service analytics. This whitepaper describes the following four scenario's to bridge the gap between the two BI forms by using CDP and CIS:

- Data virtualization as data source for data preparation
- Using data virtualization to make data sets developed with data preparation available for all users
- Using data virtualization to operationalize user-defined data sets
- Data preparation for IT specialists to develop CIS views

In addition, this whitepaper describes how these two BI forms can operate side by side in a cooperative fashion without lowering the level of self-serviceness for business users. In other words, it describes how the best of both worlds can be combined.

## 2 Phase 1 of Self-Service Analytics: IT-Driven Data Integration

Currently, with respect to data integration capabilities four phases in the evolution of *self-service analytics* can be identified; see Figure 1. This and the coming chapters describe these phases.



**Figure 1** *The four phases of self-service analytics with respect to data integration.*

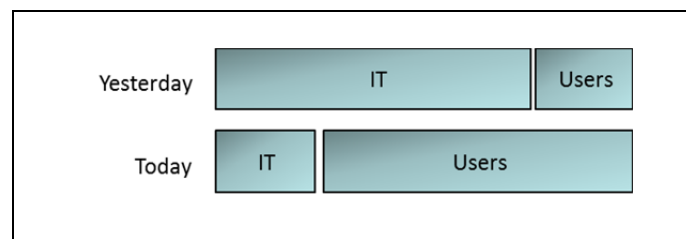
**Before Self-Service Analytics** – There was a time when IT departments were responsible for developing all the reports for every user. Users were asked what they needed, and the IT specialists would design, develop, test, and maintain the reports. IT would also make sure that the required data was available in some database, arguably a data mart or data warehouse. At that time, the only self-service aspect users had was that they could select a report from a predefined list.

Over time, users received limited *self-service* capabilities for changing and extending their reports themselves. For example, they could add dimensions, do drill-downs to study the data on a lower level of detail, and change the graphical form in which data was displayed (bar chart, line chart, or pie chart). Still, they were limited by the data made available by IT. Users could not add new data to their reports, nor could they join data from different tables in new ways. They were restricted by the data structures in the data marts and the data warehouse, which were predefined by the IT department. A change on this level would require a call to the IT department. Depending on the complexity of the request, users had to wait weeks before the new data would become available for reporting. The level of self-serviceness was severely limited by the predefined data structures.

**The Introduction of Self-Service Analytics** – With the coming of *self-service analytical tools* designed specifically for business users, such as QlikView, Tableau, and Tibco’s Spotfire, users were given the *freedom to analyze*. Also called *business or data discovery tools*, their intuitive, graphical, and easy-to-use interfaces make it possible for business managers and decision makers with minimal or no IT background to develop their own reports and to analyze the data any way they want.

Self-service analytics has become a big success, because it helps users to analyze data with the speed the business requires. The tremendous revenue growth of these vendors in the last few years confirms this popularity of self-service analytics. For example, Qlik’s annual revenues grew from 44.3<sup>1</sup> in 2006 to \$556.8<sup>2</sup> million US\$ in 2014, and Tableau grew from 20 in 2008 to 412.6<sup>3</sup> million US\$ in 2014. In both cases, these are staggering numbers, clearly showing how well the market has embraced self-service technology.

The effect of this adoption of self-service analytics by business users is that report development is increasingly performed by these users and less and less by the IT department. Gradually, IT specialists working on data warehouse and BI systems start to play the role of facilitators. They make sure the right data is available, they define semantic layers, and at that point the self-service user takes over. Because of self-service development the work is distributed differently over business users and IT specialists. Figure 2 illustrates this change of work distribution; see also the whitepaper *Streamlining Self-Service BI with Data Virtualization and a Business Directory*<sup>4</sup>.



**Figure 2** *The shift of responsibilities and work between business users and IT.*

The dominant reason for users to switch to this *do-it-yourself approach* is that IT can’t keep up anymore. Because the business is speeding up, the frequency of changes to reporting and analytics increases accordingly. The consequence is that many IT departments have an application backlog. Work on requests for new reports and changes may not start for weeks. James Kobielus<sup>5</sup> worded it as follows: “... many IT staffs face growing backlogs of information requests from increasingly frustrated end users. But self-service features ... can help. ... This frees up IT professionals from having to spend an inordinate amount of time responding to requests for new data, new views or updated report formats.”

<sup>1</sup> Andrei Pandre, *Breaking \$300K Barrier with Qlikview 11*, November 23, 2011; see <https://tableau7.wordpress.com/2011/11/23/300k-barrier/>

<sup>2</sup> Qlik, *Qlik Announces Fourth Quarter and Full Year 2014 Financial Results*, February 12, 2105; see <http://www.qlik.com/us/company/press-room/press-releases/2015/en/0212-qlik-announces-fourth-quarter-and-full-year-2014-financial-results>

<sup>3</sup> Momentum Base, *Tableau Software – Overvalued or Good Investment Opportunity?*, April 26, 2015; see <http://www.momentumbase.com/blog/tableau-software-data-overvalued-or-good-investment-opportunity>

<sup>4</sup> R.F. van der Lans, *Streamlining Self-Service BI with Data Virtualization and a Business Directory*, March 2015; see <http://www.cisco.com/web/services/enterprise-it-services/data-virtualization/documents/streamlining-data-virtualization.pdf>

<sup>5</sup> E. Horwitt, *Self-Service BI Catches On*, Computerworld, January 2011; see [http://www.computerworld.com.au/article/374318/self-service\\_bi\\_catches/](http://www.computerworld.com.au/article/374318/self-service_bi_catches/)

### 3 Phase 2 of Self-Service Analytics: Self-Service Data Integration

---

Despite the full self-service level offered for reporting and analytics by the tools, users are restricted with respect to accessing and integrating data. Users also need *freedom to integrate* data which resulted in the second phase of self-service analytics; see Figure 1.

**The Hunt for Data** – In the first phase, users are usually restricted to access the data stored in one or two data marts, seriously limiting their analytical capabilities. It enforces the following restrictions on data access:

- Data marts don't contain all the available data, only a restricted set.
- Data marts contain slightly aggregated data; users can't access data on the lowest level of detail.
- Data marts contain slightly out-of-date data that may even be 24 hours old.
- Data coming from external sources can't be accessed, only data stored in data marts.
- Data available in private or departmental files can't be accessed and combined with data marts.

This dissatisfaction has led users on a hunt, a hunt for more data. They discovered the data warehouse containing the data on the lowest level of detail, and they found the original production databases containing operational data. They also discovered the value of non-IT controlled data sources, such as social media data sites, spreadsheets on the internet developed by analyst companies containing data about the market and competitors, customer data from marketing companies, open data sources with socio-demographic data, and so on. The users' hunt for data knows no boundaries.

Currently, if users need access to other data sources, they have to ask IT to make them available by copying the data to the data mart. Such a request can take weeks or even months, while the users may need that data the same day.

**The Second Phase of Self-Service Analytics** – To reduce this data integration limitation, *data blending* (sometimes called *data wrangling*) capabilities have been added to self-service tools. The tools from, for example, Alteryx, Qlik, and Tableau, all support such features. Data blending functionality can be seen as lightweight ETL functionality. Users can develop their own logic to integrate data from different data sources. In most tools, the result of the integration process isn't stored, as is the case with ETL, but is pushed straight into the memory of the tool, making it available for reporting immediately.

Data blending and comparable functionality supported by the self-service analytical tools has started the second phase. With respect to data sources, users are not limited to access only one data mart, but have much more *freedom to integrate data*. With little or no help from IT they can integrate data from a data mart with data coming from a social media network, or data from the data warehouse with data stored in private spreadsheets.

Where phase one was primarily about freedom to report and analyze, phase two is about total freedom to integrate data sources. In this second phase data is *democratized*, meaning that all the data is made available to everyone, and not only to the IT specialists.

## 4 Phase 3 of Self-Service Analytics: Self-Service Data Preparation

---

**Understanding Data Sources** – In general, IT specialists understand the data structures of data sources, but they don't understand the data itself. The opposite applies to business users. They understand the data intuitively in a business context, because they work with it daily, but they don't understand the data structures. In other words, IT specialists think in terms of data models and specifications, whereas business users think in terms of data content.

The strength of data preparation tools, such as *Cisco Data Preparation (CDP)*, is that they're aligned with the user's style of thinking. Their starting point is the real data, not the structure or the model of the data. To be able to analyze data, users must understand the structure of the data, and CDP extracts this understanding of the data model from the data itself.

**The Complexity of Integrating Data Sources** – Is integrating data sources really easy? Is it simply a matter of clicking on the right buttons? Unfortunately, the answer is no. To integrate data sources, logic must be developed that somehow joins the data sources on the right columns and that corrects incorrect and inconsistent data. Depending on the data sources, the complexity of the logic can range from easy to highly complex. These examples illustrate the potential complexity:

- Integrating an internal customer table with an external file containing company data from, for example Dun & Bradstreet, is complex, because they don't have the same keys. The effect is that some kind of fuzzy-join must be executed based on company names and addresses. The result of such a fuzzy-join is never 100% correct, but probably close. The quality of the result is determined by the quality of the data and the smartness of the fuzzy-join.
- Integrating an operational database with a data mart can be complex, because the data structures of the production database can be cryptic and tricky. Manuals must be consulted to find out what all the columns with the coded values mean.
- Retrieving data from an operational system can be complex by itself, because data that logically belongs together is usually distributed over many tables. In a SAP database, for example, material plant data is found by combining data from the MARC, MAPR, MBEW, MDKP, and STXH tables, and a sales document item is constructed by merging data from the VBAP, VBUP, VBLB, VBKD, VBPA, and AUSP/IBIN tables.
- Integrating an internal spreadsheet containing global sales data with general market data can be difficult, because neither were developed with integration in mind.
- Integrating internal sales data with publicly-available socio-demographic data is complex when the former contains addresses and the latter is identified by XY-coordinates. To solve the problem a third file that relates XY-coordinates to addresses must be added to the join.

**The Need to Understand the Data** – Integrating data sources requires understanding of the meaning of the columns, the quality of the data, the meaning of codes, how to join columns, whether data must be transformed before joined, whether data must be corrected, and so on. Before developing the integration logic, IT specialists get that understanding by, for example, reading the available documentation on the data structures of the data sources, by talking to colleagues who have been involved in designing and



developing the data sources, and by using data profiling tools to get a feel for the characteristics of the data. This process to understand the data structures of and the data in the sources and to determine how the data can be processed, integrated, and massaged is called *data preparation*.

But how do business users do data preparation? How do they try to understand the data of internal data sources? They don't have access to colleagues who can explain it all, nor do they have time to study piles of documentation. And most of the tools that can help them to understand the data have been designed for IT specialists, not for business users. And how do they understand external data sources? Socio-demographic data, for example, may be stored in a large file in which data is organized in XML or JSON structures, or an Excel spreadsheet may be missing the column names. There may be no one to contact who can explain what all the data means.

**Self-Service Data Preparation Tools for Smart Data Integration** – Organizations can definitely increase their reporting and analytical capabilities by allowing users to use self-service reporting tools and to integrate data sources through data blending capabilities. But these tools don't help with the real process of understanding data, the data preparation work. That's the domain of *self-service data preparation tools*. These tools are designed for business analysts to assist with data preparation work by automating as much of the work as possible. *Cisco Data Preparation* (CDP) is such a tool.

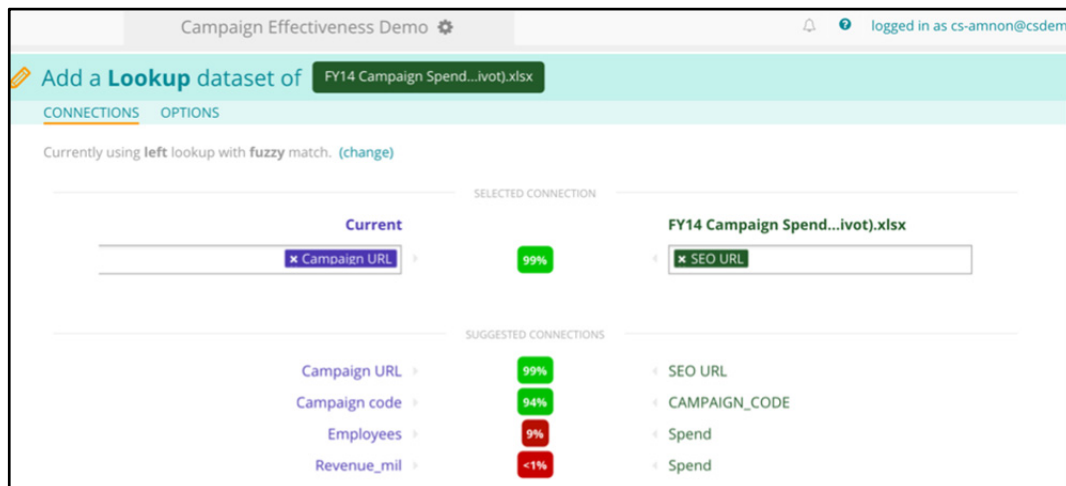
The automation underlying CDP is powerful. The user points CDP to some data source and automatically CDP analyzes the data to understand its structure and contents. Next, CDP proposes actions to join and normalize data and helps to rapidly identify and correct issues. In other words, when a data set is made available to CDP, it proposes, for example, to join on specific columns, and it helps to correct potentially incorrect values and remove outliers, and so on. Users are guided and assisted with data preparation.

With CDP, users can take any data set, prepare it, and then use it for analysis and reporting. It really enriches their analytical capabilities, and most importantly, they can do it right there and then. There is no need to call the IT department. If they locate a valuable data source, they can start to analyze it within minutes.

The rest of this section describes CDP in more detail.

**The User Interface of Cisco Data Preparation** – CDP's user interface has been designed for business users, in other words, for non-IT specialists. If users have some experience with Excel, using all CDPs features is relatively easy to learn.

Besides being intuitive, CDP's user interface plays an important role in the process of data preparation. For example, coloring specific values in columns is used to identify groups of records, or a simple listing of potential joins is used to show what the best way is to integrate two data sets. To illustrate, Figure 3 shows CDP's proposal for joining two data sets. The two green boxes show the two preferred joins and the red boxes the weakest joins. Joining the column Campaign\_URL from the first data set with the column SQL\_URL from the second is rated as best by CDP, because 99% of the values appears in both columns. The second best join is the one on the columns called Campaign code, because only 94% of the values appears in both columns. The only thing the user must do to integrate the data sets is to accept one of these proposals by clicking one of these buttons.



**Figure 3** Proposed joins by CDP sorted by effectiveness.

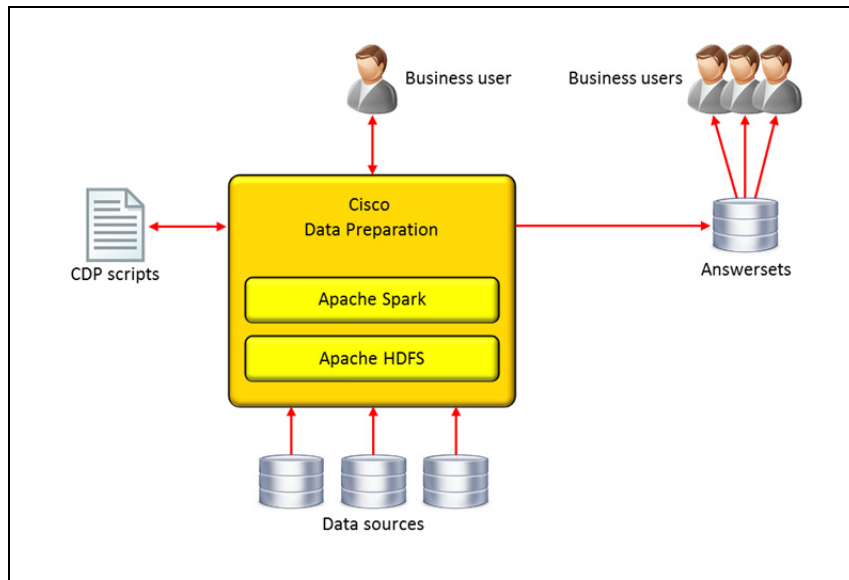
In Excel, when formulas are added to a spreadsheet or if values are changed, the entire spreadsheet is recalculated instantaneously and the result is visualized in the spreadsheet. It’s all happening real-time. CDP uses the same approach, for example, when a user decides to remove some outlier values (as proposed by CDP), they are removed from the data set right away. It’s the same with joins, when a user picks one of the proposed joins, the real data making up the result of the join is presented. Also, when such operations have to be undone, those operations are executed instantaneously as well.

Overall, CDP’s user interface has been designed to support a highly interactive process whereby tool and business user together transform step by step the source data to a form that’s right for analysis. During this process, users can freely try out alternatives, undo previous decisions, add more data, and so on.

**Machine-Learning Algorithms** – To analyze the data and to come up with proposals automatically for preparing the data, CDP uses *machine-learning algorithms*. Some of the operations supported by CDP are automatic join detection and data normalization. Users don’t have to analyze what the best join is. It’s done automatically and that saves them a lot of time. Also, it helps lesser-skilled analysts to perform like highly-skilled data scientists.

**The Architecture of Cisco Data Preparation** – Figure 4 shows the overall architecture of CDP. Internally, CDP’s architecture is developed on Apache Hadoop’s modern-day distributed storage technology and Apache Spark’s highly scalable processing technology.

Currently in CDP, data for data preparation must be stored in HDFS files. This means that data stored in other data storage mechanisms must be copied to HDFS. There are several reasons why data is copied: to complete lineage, to guarantee repeatability, and to homogenize local, 3rd-party, Hadoop and other data source formats. Also, the data structures used by CDP when storing the data in Hadoop are optimized for data preparation to speed up its own processing.



**Figure 4** The overall architecture of CDP which is based on Apache Hadoop's distributed storage technology and Apache Spark's highly scalable processing technology.

When users apply operations to data sets, CDP automatically documents them in *scripts*. It's up to the users whether they want to see these scripts or not, but it's not necessary for the data preparation task. These scripts can be shared with other users and can be re-executed later on.

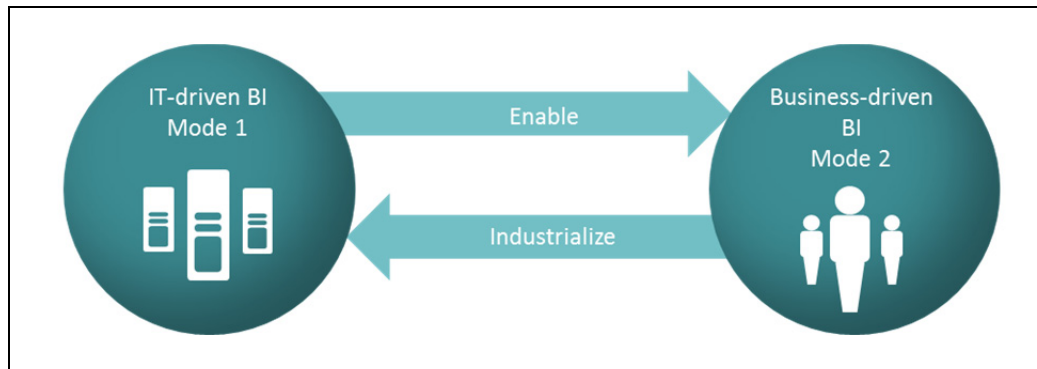
When users think they're ready, they can export the result to a HDFS file called an *Answerset*. CDP can publish answersets in Apache HCatalog so that other tools can retrieve the data structure to understand the contents of the answerset. For example, Hadoop's SQL-based interfaces, such as Apache Hive and Impala, can access these answersets when defined in HCatalog. And via these interfaces almost any reporting and analytical tool can access the data.

Note that in no circumstances does CDP change the original source data. What CDP shows in the user interface is the data after all the preparation operations have been applied, with this result kept in memory. The data in the original source is not updated. This approach is very similar to developing a view in CIS. The view does not change the data source either, but allows users to look at the data after the operations have been applied. Exporting data to an answerset with CDP is like materializing the contents of a view as a physical table with CIS.

## 5 Phase 4 of Self-Service Analytics: Cooperative Data Integration

**IT-driven BI versus Business-driven BI** – Besides self-service analytics, many other forms of reporting and analytics exist, such as standard reporting, embedded analytics, ad-hoc reporting, and mobile BI. Commonly, organizations attempt to have one BI environment to support all these different forms. BI forms can be divided in two categories: *IT-driven BI* and *Business-driven BI*; see Figure 5.

With *IT-driven BI* all the reports are designed and developed (partially or fully) by IT and must be tested, governed, and audited. Typically, these reports must be reproducible years later. Some of them must be distributed to regulators, customers, or suppliers. Typical terms associated with this form of BI are: industrial scale, specification-driven process, professionally tested, SLA metrics, governance, and control.



**Figure 5** *Two modes of operation: IT-driven BI and Business-driven BI.*

*Business-driven BI* relates to all the reports designed and developed by business users and business analysts. Users can also change their own reports at any time. Self-service BI, investigative analytics, data discovery, and data science are all forms of business-driven BI. Typical characteristics of this second form of BI are: self-service, agile, non-governed, non-auditable, investigative, one-time usage, no or minimal sharing.

This categorization of BI forms is in line with Gartner's *BI-modal* concept<sup>6</sup>. IT-driven BI corresponds with Gartner's *Mode 1* and business-driven BI with *Mode 2*. Both terminologies are used interchangeably in this whitepaper.

**The Fourth Phase of Self-Service Analytics is About Cooperation** – The first three phases all aim at making self-service a better experience and improve productivity of the business users. In Phase 4, the focus is to let the self-service environment work cooperatively with IT-driven BI.

Supporting both BI forms or modes can be a technological challenge. It's not easy to develop a data warehouse that is tuned to support fixed forms of reporting for which tight performance SLA's apply plus an ad-hoc, self-service environment in which no predictions can be made with respect to workload. Technically, the easiest solution to support both forms is by developing two separate environments. This may be easiest, but has some severe drawbacks: consistency of report results from the two environments is hard to guarantee, high storage costs, double management costs, and duplication of report development work is difficult to avoid, and so on.

That organizations have to support both forms is a fact. In a Gartner survey<sup>7</sup>, 45 percent of CIOs state they currently have a Mode 2 operation and by 2017 they predict that 75 percent of IT organizations will have a BI-modal capability, meaning they support both modes.

Besides being able to operate the two forms of BI, it's important to bridge the gap between the two; see Figure 5. Especially, because reports may be developed in Mode 2 initially and have to be moved to Mode 1 afterwards. For example, a marketing manager may have developed a report that analyzes in some creative way data from an internal source with an external source. Such a report may be useful for colleagues as well, and maybe they want a refresh of that report every week. This report may have been

<sup>6</sup> M. Mesaglio and S. Mingay, *Bimodal IT: How to Be Digitally Agile Without Making a Mess*, July 2014; see <https://www.gartner.com/doc/2798217/bimodal-it-digitally-agile-making>

<sup>7</sup> P. Sondergaard, *Bimodal Business Leads to Bimodal IT*, December 2014; see <http://blogs.gartner.com/peter-sondergaard/bimodal-business-leads-to-bimodal-it/>

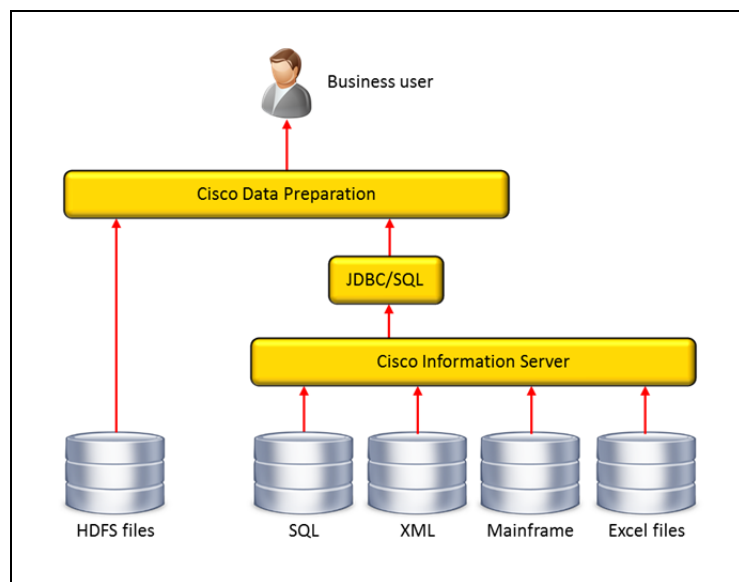
developed in Mode 2, but to enlarge its audience, it must be migrated to Mode 1. This is referred to as *industrialization or operationalization* of a Mode 2 report.

Another way to bridge the gap is when metadata and reporting specifications developed in Mode 1 are shared by users operating in Mode 2. For example, the integration of data sources may be so complex that it's developed by IT and then handed over to the Mode 2 users for more interactive forms of reporting. In this situation, IT *enables* Mode 2 usage.

**The Solution is Combining Data Preparation with Data Virtualization** – To support both BI forms and to bridge the gap between the two modes, and to fully enable the freedom to integrate cooperatively, two technologies are required: *data preparation* and *data virtualization*. The following sections describe four scenarios to bridge the gap between Modes 1 and 2 by using Cisco's Data Preparation product (CDP) and Cisco's Data Virtualization product called CIS. In addition, they describe how these two modes can operate side by side in an integrated fashion without lowering the level of self-serviceness for Mode 2. In other words, it describes how the best of both worlds can be combined.

## 6 Scenario 1: Data Virtualization as Data Source for Data Preparation

**CIS as Data Source for Data Preparation** – CDP offers access to many different data sources, including Apache HDFS files, spreadsheets, text files, XML documents, and those that can be accessed through a JDBC interface. CIS offers a JDBC interface that can be accessed by CDP. Every view defined in CIS can be used by CDP as a data source. With CIS, views can be defined on a large set of data sources, including almost any SQL database server, Excel files, flat files, data on mainframes, OLAP cubes, JSON documents, and web services. Currently, many of these data sources cannot be accessed by CDP directly, but through CIS. CIS extends the breadth of data sources that can be analyzed and prepared by business users with CDP; see Figure 6.



**Figure 6** CIS extends the list of data sources users of CDP can access.

When CDP accesses a CIS view, CIS determines the content of the view by accessing the data sources or a cache. The result is then transmitted to CDP that loads it into a HDFS file for further processing. So, when

CDP analyzes the data, CIS is not involved anymore, nor are the original data sources. CIS is only doing some processing at the beginning to get the data. Most of the CDP processing is done standalone.

As an example, Figure 7 shows a screenshot of CDP in which it accesses a source called Orders, which is a view defined in CIS. The bottom half of the screenshot shows the CDP library and contains a list of some of the data sets users can access. The technical connection string to get to the Orders data is shown at the top of the screenshot.

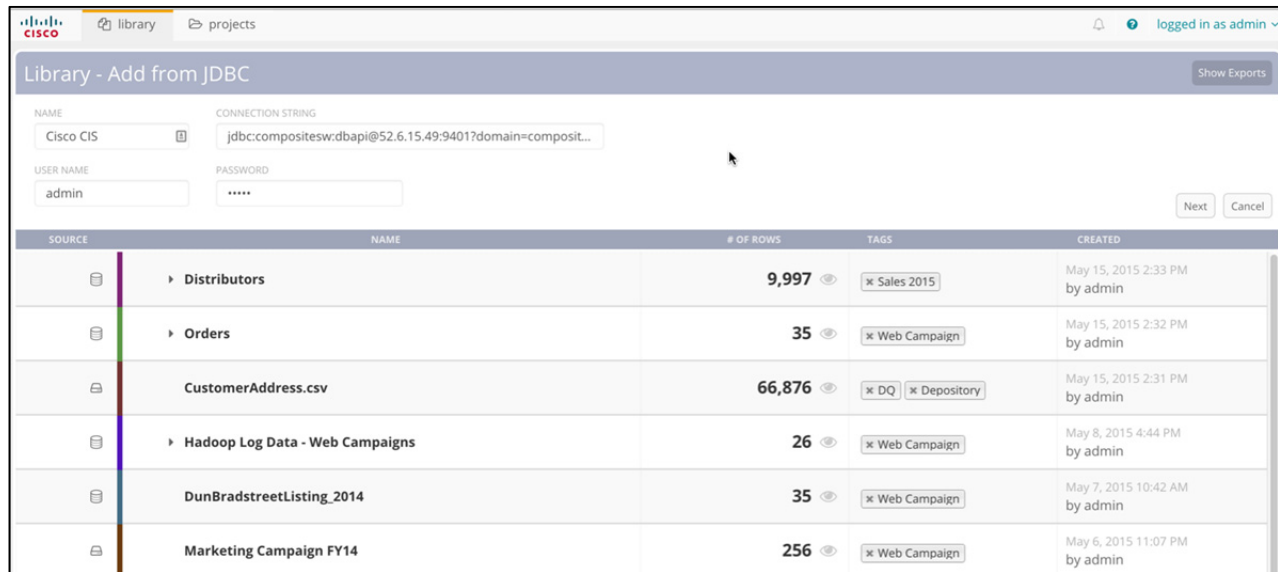


Figure 7 Using CDP to connect to the CIS view called Orders.

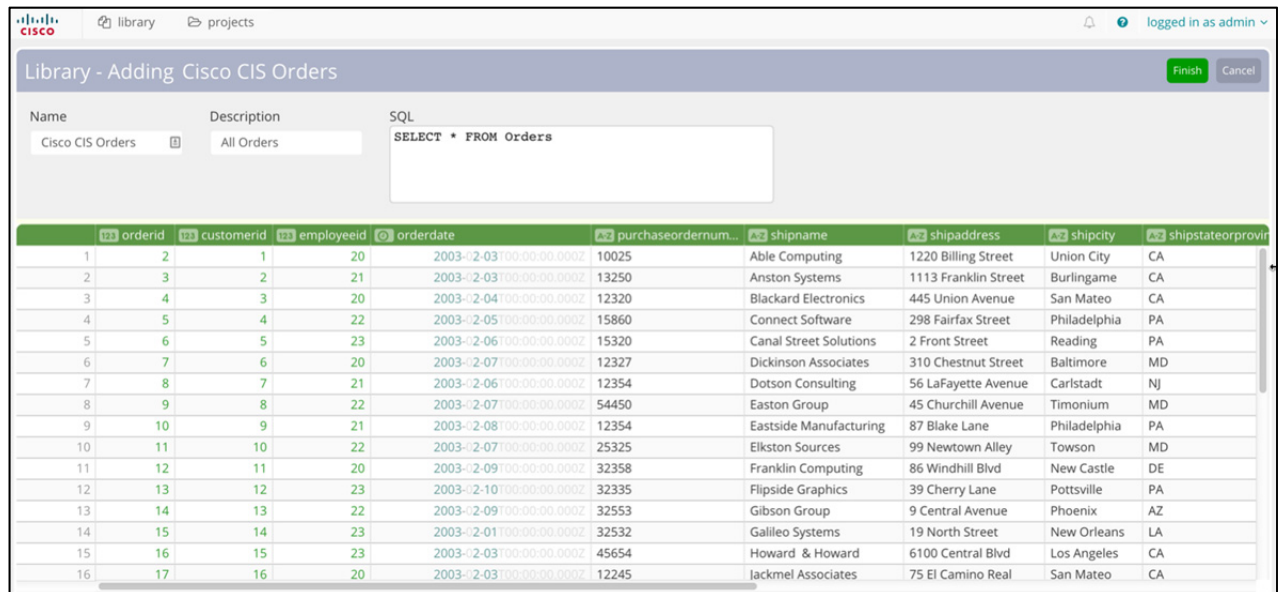


Figure 8 CDP showing the contents of the Orders view as defined in CIS.

Figure 8 shows a SQL statement that is executed on the Orders view in CIS plus the result of the query. Note that neither the users, nor CDP know the original data structure. For example, it can be data stored in an XML document that has been flattened, data coming from a spreadsheet that has been integrated

with data from a web service, or data stored in a VSAM file on a mainframe. CIS hides all these (irrelevant) technical aspects for the business users.

**Benefits of Using CIS as a Data Source** – To let CDP access CIS views offers several benefits:

- **Increase breadth of available data sources:** If data is stored in a data source that isn't supported by CDP but by CIS, that data becomes available quickly by defining a simple one-to-one view on that data source and allowing CDP users to access the view. For example, some data may only be available through invocation of a web service through a REST or SOAP interface. With CIS a view can be defined on such a web service. In this view definition the non-relational data can be made available via JDBC and SQL.
- **Simplify data access for business users:** Many business users are not technical enough to work with some data sources. For example, if data is stored in an XML-document with a hierarchical data structure, an IT specialist may be needed to transform it in a flat data structure that CDP can use. By defining a view in CIS on this XML document, access to this data becomes much easier for business users.
- **Simplify integration of data:** Even with CDP, integration of some data sources may be complex. Especially when data is highly coded and data has been distributed over many tables, it can be hard for users and CDP to analyze the data. In this case, IT specialists can first define CIS views to simplify the data structure and transform cryptic and coded data to more meaningful values. It is still up to the users to determine how they want to prepare the data for reporting and analytics.
- **Access to critical enterprise systems:** Every mission-critical enterprise system contains valuable data for business users. Unfortunately, in many organizations users are forbidden to access these data sources using their data preparation tools, because it can interfere with their operation. In such a situation, the IT department responsible for these systems will probably recommend users to analyze the data stored in the data marts. But, as indicated, this data may be slightly aggregated, filtered, and probably even slightly out of date. An alternative solution is that CIS views are defined on the tables in the enterprise systems and subsequently cached. Next, these views are made available to the business users for use with CDP. This way, analyzing the data doesn't interfere with the existing workload on the enterprise systems. Plus, getting a new and more up-to-date copy of the enterprise data is purely a matter of a simple refresh of the views.
- **Avoid reinvention of the wheel:** Many users may access the same data sources. If they all have to prepare the same data from scratch, eventually they will all perform the same activities, and all will develop similar logic to transform the source data in a more meaningful structure and form. In other words, the wheel is invented over and over again, which is a waste of valuable resources. This can be avoided by defining some of the general rules that apply to all the users in CIS views, and allowing all the users to access the same view definitions; see also Scenario 3 in Section 8.
- **Access to a virtual data lake:** To make data available via a *data lake*<sup>8</sup>, it must first be copied to the data lake. For many data sources, copying data makes a lot of sense, but not for all. Especially really big data files may be too big to copy. Too big means that the storage costs are too high, the copying takes too long, or the entire copying process is too expensive. In this case a *virtual data*

---

<sup>8</sup> Techopedia, *Data Lake*; see <http://www.techopedia.com/definition/30172/data-lake>

*lake*, sometimes called a *hybrid data lake*, is more practical. With a virtual data lake all the data is accessible via CIS. In other words, CIS pretends a large data lake exists. Some of the data being accessed is really stored in a physical data lake and accessed directly by CIS, whereas some of the data is retrieved by CIS when users ask for it. That data is not stored in the physical data lake. A virtual data lake can be more cost-effective.

**The Business Directory** – Cisco’s *Business Directory* (CBD) helps in documenting and searching all data sources and data sets. The whitepaper *Streamlining Self-Service BI with Data Virtualization and a Business Directory*<sup>4</sup> describes CBD in detail and how CBD and CIS can operate together.

An organization may have hundreds of data sources and each can have hundreds of tables with data. In addition, as indicated, more and more external data sources are available. So, before users can point CDP to a specific source, they must first find the right one. For example, if users are interested in sales data, which tables contain the right data for their reports? There may be tens of tables that contain sales and sales-related data. It’s not unimaginable that users can’t see the forest for the trees anymore. To paraphrase John Naisbitt’s<sup>9</sup> famous quote: “Users are drowning in data sources, but they’re starving for information.”

Users need a mechanism where they can enter search-terms, such as sales or factory spill, and all the tables in all the data sources that contain sales or factory spill data are shown. Users need a Google-like mechanism to locate the right data. This is where CBD comes in. CBD allows users to introduce business terms to describe tables and data sources in non-technical terms, to organize and categorize the terms, and to search the terms. All this business information allows users to find the right data needed for reporting and analytics. It creates a *business context* for all the available data. In a nutshell, CDP opens up the data lake to self-service business users. Even undocumented data for which all the metadata is missing, can be derived and can be used afterwards for reporting and analytics. An in-depth description of CBD can be found in the whitepaper *Streamlining Self-Service BI with Data Virtualization and a Business Directory*<sup>4</sup>.

To speed up self-service analytics, the general recommendation is to provide access to all the data sources (including all the CIS views) via CBD. It allows users to locate the data they need much faster.

## 7 Scenario 2: Using Data Virtualization to Make Answersets Developed with Data Preparation Available for Everyone

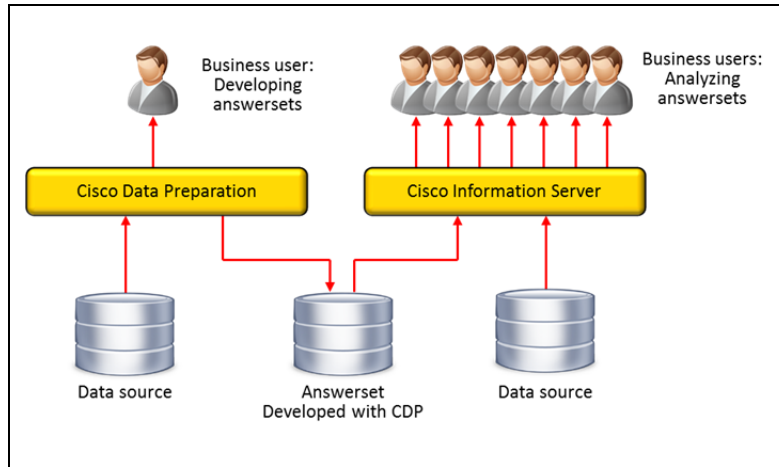
---

After preparing data, an answerset can be created in which all the operations that the business user identified, have been applied to the data. This answerset can be accessed or imported by any reporting or analytical tool. Currently, CDP can only export the answersets as Hadoop HDFS files. When they are made available, all the users have access to all the data in these files. An alternative approach would be to make the answersets available to the users via CIS as depicted in Figure 9. This approach makes the data available for a much larger audience.

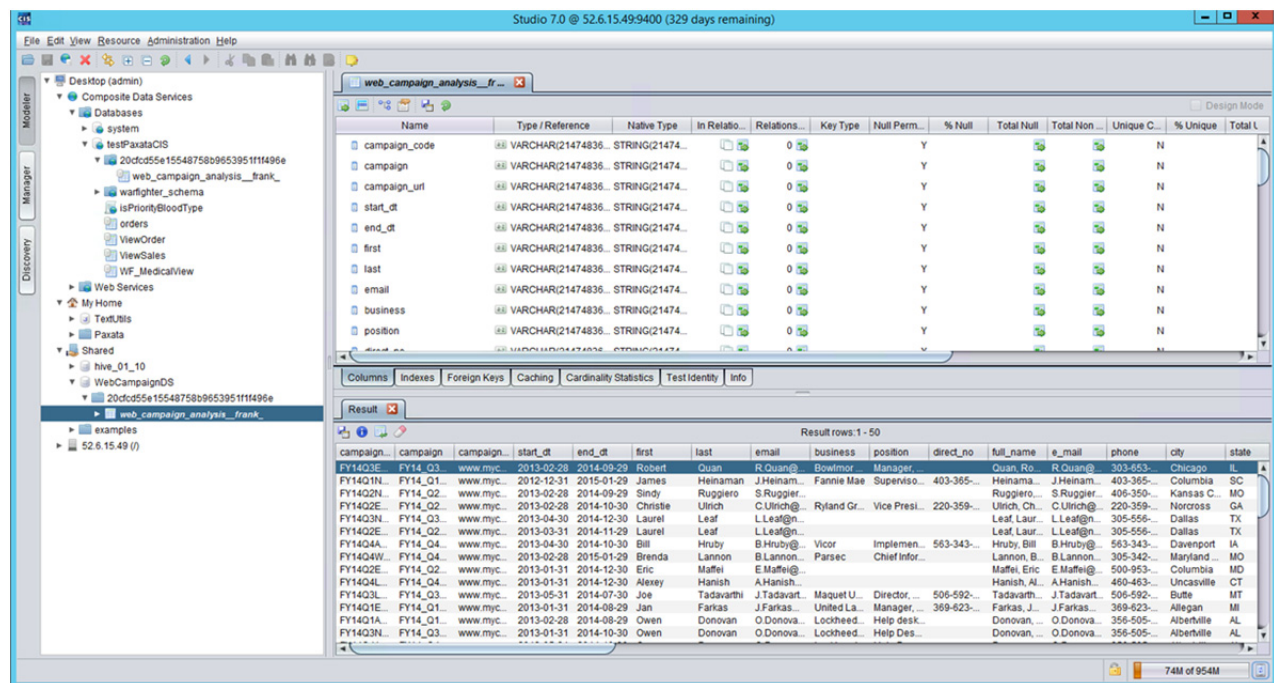
---

<sup>9</sup> John Naisbitt, *Megatrends; Ten New Directions Transforming Our Lives*, Warner Books, first edition, 1982.





**Figure 9** Using CIS to make answersets developed with CDP available to a larger audience.



**Figure 10** An answerset developed with CDP is imported in CIS using Apache Hive.

**Making CDP Files Available for CIS** – The process to make CDP answersets available through CIS is a relatively straightforward process:

- Use CDP to export the answerset to an HDFS file.
- Use CDP to document the answerset in Apache HCatalog. The effect is that the data structure (column names, data types, and so on) of the answerset is documented in HCatalog; see also Section 4.
- Log in with CIS and import the HDFS file as a table through the JDBC interface for Apache Hive. As an example, Figure 10 shows how a table called `web_campaign_analysis__frank` is imported. It also displays some rows of that table so that developers can see whether the right table has been selected.

- Define security rules for the users and user groups that are allowed to access the data. Define the view as read-only, because allowing changes to this file makes no sense, because it's derived by CDP from other files.
- Publish the developed view.
- Define and document the view in the Cisco Business Directory. This helps users to find the data more easily.

Remarks: Note that an answerset developed with CDP is not updated when the source is updated. As indicated, the answerset is derived from other data sources by CDP. To update the contents, the script developed with CDP must be re-executed and the result must be exported again. The newly created answerset must replace the existing one.

**Benefits of Making CDP Results Available Via CIS** – There are several benefits of making an answerset developed with CDP available via CIS:

- **Bigger audience:** When data created with CDP becomes available through CIS, it can be accessed by almost any tool, from a simple reporting tool to the most advanced analytical tool. So, the data can then be used by a wide range of users within the organization.
- **Large set of APIs:** When an answerset is available via CIS, it can be accessed through a large set of languages and APIs. And if required, the data can also be presented in another form. For example, the flat data structure created by CDP can be turned into a more hierarchical structure based on XML and JSON. Or, the data can be made available via an MDX interface, which is what specific analytical tools require.
- **Data aggregation:** An answerset may contain a level of data that is too detailed for some users. Aggregating the data by defining an extra view on top of the original view, allows those users to see the same data in an aggregated form without the need to create an extra file that contains the aggregated data. By defining multiple views, different users can see the same data with different aggregations.
- **Data filtering:** What applies for aggregation, applies for filtering as well. Different views can be defined that show different subsets of columns and records from the original answerset.
- **Combining with other data:** An answerset created with CDP and made available via CIS can be enriched with other data sources. For example, one view can be defined that joins the CDP data with some other views already defined in CIS.
- **Data security:** Data used by BI users has to be protected against unauthorized usage, because it's usually business critical and business confidentiality. With CDP, users can be restricted to access particular answersets. This mechanism only works on full answersets. CIS offers more fine-grained data security mechanisms. So, by making the answersets available via CIS and by using CIS' security mechanisms, subsets of answersets can be made available to users. In this case, the answerset itself must be protected on the file system level as well; CIS must become the only "user" of the answerset making sure all access to the data is handled through CIS.

## 8 Scenario 3: Using Data Virtualization to Operationalize User-Defined Data Sets

---

**Why Operationalizing CDP Scripts?** – Answersets developed with CDP are very valuable to users who developed the CDP scripts. But despite their potential value, for some users they may not be what they need:

- First, answersets created by CDP are like *snapshots* of the data. Their contents are not automatically updated when the source files are updated. So, eventually they may contain data that is too “old” for some users. To update an answerset, the script must be executed by CDP again.
- Second, some users may require that the process of applying the scripts is governed and audited using more traditional IT-based techniques.
- Third, running reports on the answersets means accessing the data using Apache Hive or some other SQL-on-Hadoop engine. It may be that a more scalable and faster implementation is required.

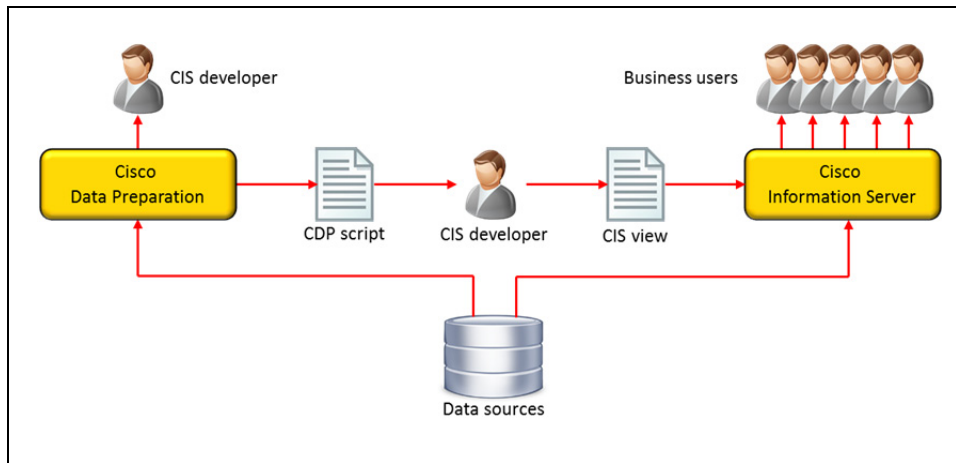
For all these situations it’s recommended to *operationalize* CDP scripts using CIS: logic developed in CDP is migrated to CIS. In other words, CDP scripts are translated to CIS SQL-based views. When this is done, users accessing the CIS views see up-to-date data, there is no need to schedule updates of the data, and CIS can be used to offer the required performance and scalability level.

In fact, when migrating the logic to CIS, development of the CDP scripts can be seen as a *prototyping* activity to quickly analyze all the operations that must be applied to the data to make it useful for reporting and analytics.

Note that the key difference between this and the second scenario is that with the former the implementation of the CDP script is removed and replaced by CIS views. With the third scenario the CDP script is a prototype and the CIS view is the end result, while with the second scenario the answerset is the end result, and there is no real intention to repeatedly update its contents.

**Reverse Engineering the CDP Script** – No tool exists that is capable of automatically transforming CDP scripts to CIS views. The operationalization process is for a large part a manual process; see Figure 11. However, many of the specifications making up a CDP script can be implemented one-to-one. For example, if a filter is implemented to remove specific rows from the source data, this specification can be implemented as a WHERE clause. But not all the operations can be translated that easily. For example, if only a complex fuzzy-join is possible to link two tables together, this join cannot be implemented as a simple equi-join in CIS. The specialist has to evaluate what to do.

It’s recommended to provide access to each view defined in CIS in the Cisco Business Directory so that users can search for them more easily.



**Figure 11** *Operationalizing CDP results using CIS.*

**Benefits of Operationalizing CDP Scripts** – Some of the benefits that apply to scenario 2 apply to scenario 3 as well:

- **Bigger audience:** When data is available through CIS, it can be accessed by almost any tool and thus many users.
- **Data security:** CIS offers more fine-grained data security mechanisms than CDP.

And there are benefits specific to Scenario 3:

- **Up-to-date data:** CIS views can show up-to-date contents.
- **Consistent results:** By running the same logic every time a view is accessed, all reports, including the ones developed with different tools, are consistent. For example, when two users access the view, one using Excel and the other Tableau, the values they see are identical.
- **Improved performance and scalability:** The full power of CIS can be used to make sure access to the views is as fast as possible. The advanced join optimization techniques of CIS, the caching mechanisms, and the server-based architecture can all be used to offer the right level of performance and scalability.
- **Governed and audited:** Some of the reports operating on the data may have to be governable and auditable using traditional IT governance techniques, such as where-used, lineage, and use logging. CIS has the features to make that possible.
- **SLA:** When performance, availability, and scalability levels must adhere to specific SLAs (Service Level Agreement), CIS has more options available to meet those SLAs.

## 9 Scenario 4: Data Preparation for IT Specialists to Develop CIS Views

In the first three scenarios it's the business user using CDP. In the fourth scenario for combining data virtualization and data preparation, it's the CIS developer using CDP. In a way this fourth scenario shows how CDP can be used in Mode 1 (see Figure 5) to increase the productivity of CIS developers. In Chapter 5 this is referred to as enabling Mode 1.

**Data Virtualization is About Agility** – CIS and other data virtualization servers are commonly acquired by customers to improve the productivity of developers and to increase the agility of BI systems; see also the whitepaper *Data Virtualization for Business Intelligence Agility*<sup>10</sup>. Productivity increases because the integration of data sources can be implemented in a fraction of the time in which the same can be done with ETL tools. This is not because ETL tools are complex, but because ETL-based solutions require extra databases to be developed and maintained. The whitepaper *Migrating to Virtual Data Marts using Data Virtualization*<sup>11</sup> compares in Chapter 3 all the activities to design, develop, operate, and maintain ETL solutions versus those for data virtualization solutions.

**Improving Productivity with Data Preparation** – But what if IT specialists don't have a deep understanding of the data and data structures of the data sources they have to integrate with CIS? Or what if there are problems with the data quality? As CDP doesn't require this understanding to get started, IT specialists can use CDP to easily build up this understanding. In other words, in the same way that CDP makes it easy for business users to prepare the data, it can assist IT specialists when integrating data sources with CIS; see Figure 12. In other words, CDP can be a powerful and productivity-improving instrument for IT specialists when designing and defining views in CIS. It makes CIS developers even more agile.

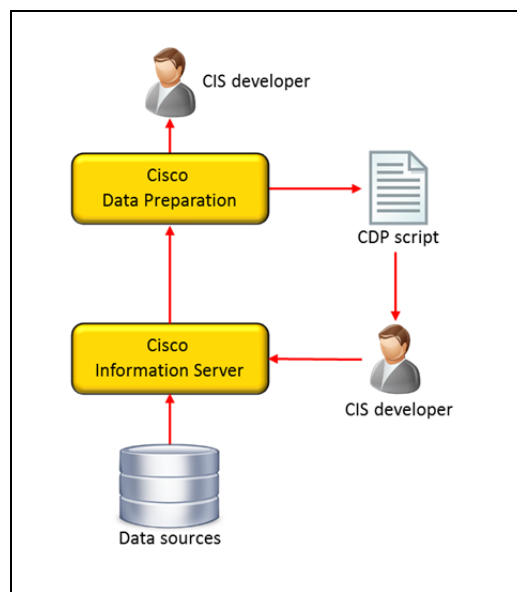


Figure 12 CIS developers using CDP to assist with data preparation.

<sup>10</sup> R.F. van der Lans, *Data Virtualization for Business Intelligence Agility*, February 2012; see [http://purl.manticoretechnology.com/MTC\\_Common/mtcURLSrv.aspx?ID=12917&Key=FE72CA6B-C6D6-4DA1-91D6-5CDE20B85E33&URLID=17966](http://purl.manticoretechnology.com/MTC_Common/mtcURLSrv.aspx?ID=12917&Key=FE72CA6B-C6D6-4DA1-91D6-5CDE20B85E33&URLID=17966)

<sup>11</sup> R.F. van der Lans, *Migrating to Virtual Data Marts using Data Virtualization*, January 2015; see [http://www.cisco.com/web/services/enterprise-it-services/data-virtualization/documents/Whitepaper\\_Cisco\\_VirtualDM.pdf](http://www.cisco.com/web/services/enterprise-it-services/data-virtualization/documents/Whitepaper_Cisco_VirtualDM.pdf)

**Benefits of Data Preparation for CIS Developers** – There are several benefits of allowing developers to use CDP when designing their virtual environment:

- **Improved productivity:** Until now, uncovering data-related issues has been a manual activity for CIS specialists. With CDP this activity can be speeded up, because potential problems with the data can be discovered very early in the project and not after all the views have been defined and after users have started to develop their reports. Again, this is because CDP “hunts” for problems that CIS developers may miss completely.
- **Unfamiliarity with data sources:** Especially if CIS developers are unfamiliar with some data sources, CDP can speed up the process of understanding all the peculiarities and characteristics of the data. For example, incorrect data, the best way to integrate tables, and the need to de-duplicate data, can all easily be identified with CDP. It’s as if the specialists use CDP as a reverse engineering tool.
- **Improve data quality:** When CDP assists CIS developers by applying its machine-learning algorithms, data quality issues may become more easily visible than when the data preparation activity is executed manually. This helps to improve data quality and thereby the quality of the reports accessing the CIS views.

## About the Author Rick F. van der Lans

---

Rick F. van der Lans is an independent analyst, consultant, author, and lecturer specializing in data warehousing, business intelligence, database technology, and data virtualization. He works for R20/Consultancy ([www.r20.nl](http://www.r20.nl)), a consultancy company he founded in 1987.

Rick is chairman of the annual European Enterprise Data and Business Intelligence Conference (organized annually in London). He writes for [Techtarget.com](http://Techtarget.com)<sup>12</sup>, [B-eye-Network.com](http://B-eye-Network.com)<sup>13</sup> and other websites. He introduced the business intelligence architecture called the *Data Delivery Platform* in 2009 in a number of articles<sup>14</sup> all published at [BeyeNetwork.com](http://BeyeNetwork.com). The Data Delivery Platform is an architecture based on data virtualization.

He has written several books on SQL. Published in 1987, his popular *Introduction to SQL*<sup>15</sup> was the first English book on the market devoted entirely to SQL. After more than twenty-five years, this book is still being sold, and has been translated in several languages, including Chinese, German, and Italian. His latest book<sup>16</sup> *Data Virtualization for Business Intelligence Systems* was published in 2012.

For more information please visit [www.r20.nl](http://www.r20.nl), or email to [rick@r20.nl](mailto:rick@r20.nl). You can also get in touch with him via LinkedIn and via Twitter @Rick\_vanderlans.

## About Cisco Systems, Inc.

---

Cisco is the worldwide leader in helping companies seize the opportunities of tomorrow by delivering the amazing results that come from connecting the previously unconnected. Cisco is empowering countries, cities, industries and businesses around the globe to move faster in order to keep pace with digital transformation and the move to the Internet of Everything (IoE). Cisco Information Server is agile data virtualization software that makes it easy for companies to access business data across the network as if it were in a single place. Cisco Data Preparation is self-service software that makes it easy for non-technical business analysts to gather, explore, cleanse, combine and enrich the data that fuels analytics. For more information, please visit [www.cisco.com/go/datavirtualization](http://www.cisco.com/go/datavirtualization).

---

<sup>12</sup> See <http://www.techtarget.com/contributor/Rick-Van-Der-Lans>

<sup>13</sup> See <http://www.b-eye-network.com/channels/5087/articles/>

<sup>14</sup> See <http://www.b-eye-network.com/channels/5087/view/12495>

<sup>15</sup> R.F. van der Lans, *Introduction to SQL; Mastering the Relational Database Language*, fourth edition, Addison-Wesley, 2007.

<sup>16</sup> R.F. van der Lans, *Data Virtualization for Business Intelligence Systems*, Morgan Kaufmann Publishers, 2012.