This chapter provides an introduction to the tools SAP offers to help provision data for SAP HANA. It begins with a look into what types of tools you have to choose from; then, it dives a little deeper into what sets each tool apart.

“Introduction”

Contents

Index

The Authors

Megan Cundiff, Vernon Gomes, Russell Lamb, Don Loden, Vinay Suneja

Data Provisioning for SAP HANA

352 Pages, 2018, $79.95
ISBN 978-1-4932-1671-0

www.sap-press.com/4588
Chapter 1

Introduction

When it comes to data provisioning, most companies have to work with the data and the tools they have. We hope this book will help you make the right choices as you navigate provisioning data to SAP HANA.

If you deal with data, whether large or small, you’ll probably ask yourself at some point, “How can I get this file/table/extract/feed into SAP HANA?”

If you haven’t heard this question a hundred times already, you will soon. Project managers schedule meetings on this question; analysts ping every IT contact they know searching for a quick answer. When asking an SAP HANA consultant, the answers might border on endless. The alphabet soup of solutions and tool names can be confusing even to seasoned SAP users.

Whether you’re an IT executive or a developer, your customers are probably asking this question, and your goal should be to provide a simple answer, which will require at least a cursory understanding of the available tools, an inventory of the tools currently available to you, and a methodology for determining the best solution for your users’ circumstances. This book aims to strengthen you in all three areas, so that you can quickly and confidently leverage SAP HANA’s in-memory computing to support your organization. First, let’s look into what types of tools we have to choose from; then, we’ll dive a little deeper into what sets each tool apart.

1.1 What Are the Tools for Provisioning Data?

The hardest part is usually getting started. We’ll cover six tools in depth in this book, but we can group them into three categories to help you quickly decide where to focus your efforts: ETL (extract, transform, and load); cleansing; and replication. Let’s briefly define each category and see how the six tools fall into each category; then, we can dive a little deeper into what separates these tools from others in the market.
Often, to be clear and concise, the meticulous grouping of functionalities into acronyms can have the opposite effect. Suddenly, rather than saying, “You can use SAP HANA’s built-in ETL tool,” you might end up saying, “You can use SDI via SDA and a Data Provisioning Agent server.” Despite meaning the same thing, the latter statement can easily result in hours researching and making lists of pros and cons.

But, ultimately, each tool has its place, and in this section, we’ll clarify the overarching use case for each. First, SAP HANA smart data integration (SDI) is a tool primarily focused on getting your SAP HANA system up and running as quickly as possible by being bundled with the platform natively. Next, SAP Data Services is designed to create a common language across your organization, which may or may not include SAP HANA, and facilitate data movements. Third, SAP Agile Data Preparation peeks behind the curtain a bit to allow business users build their own joins and lookups on source data. Finally, the SAP Landscape Transformation Replication Server (SAP LT Replication Server) is a tool that you can use to quickly put SAP HANA to work and start querying massive amounts of SAP data.

Separating the tools into these broader categories hopefully points to a larger theme in this book, which is that no one tool can do it all, all the time. More often than not, a combination of these tools is required to support a large organization with data spread out across multiple SAP and non-SAP systems.

We’ll look at each tool independently to understand its strengths and weaknesses and its place in the IT landscape. If you already know which tools you plan to use, skip to the specific chapter for the nuts and bolts of utilizing the tool in your provisioning strategy.

### 1.1.1 Extract, Transform, and Load

ETL products enable you to manipulate your data before loading the data into SAP HANA. By offering standardization and reproducible data enhancements, ETL tools can greatly improve analyst productivity by removing repetitive tasks from the daily workload. If a user mentions they need to download or export the data into Excel so that the data can be “massaged” or “cleaned up” before uploading, an ETL tool can be inserted into the process to automate those tasks, thus allowing your analysts to focus on analysis. When provisioning SAP HANA, if one of your users says, “I have a file,” the first question you should ask is “How do you get this file?” The answer will help you decide between the two provisioning tools found in this group, as follows:

- SAP Data Services
- SAP HANA smart data integration (SDI)

### SAP Data Services

SAP Data Services is a one-stop-ETL-shop for SAP data integration. Other ETL tools exists, of course, such as Informatics, SSIS, and open source options such as Pentaho, but for multisystem integration in a mixed landscape that includes any amount SAP software, SAP Data Services is the ETL tool of choice because of ability to natively access SAP programs and its change data capture options. However, using SAP is not a prerequisite for using SAP Data Services.

SAP Data Services’ primary function is to provide a layer across all data storage devices in your organization, both on-premise and in the cloud. SAP Data Services includes eight customized ODBC adapters, can utilize JDBC connections, parse Hadoop file stores, import web services for software-as-a-service (SaaS) integrations, open FTP and SFTP file locations, connect to Samba and Windows shares, and in a pinch even leverage Windows and Unix shell commands and custom Python scripts. In terms of data storage, SAP Data Services levels the playing field by providing a single syntax to interface with all these storage options. Let’s look at a few examples to expand on this topic from a developer’s point of view.

### The Tool of Many Names

Another common name for SAP Data Services is the “Data Integrator (DI)” or the “SAP BusinessObjects Data Integrator (BODI),” which is used to refer to the same tool, minus the data quality transforms used for data cleansing. This licensing difference is often overlooked by developers who may simply refer to the tool as SAP Data Services.

For anyone who has worked with any type of data, SQL (Structured Query Language) is not a new term. But, too often, many forget that not all SQL is created equal. Every database has its own unique features and solutions for certain tasks and, thus, also unique syntax requirements. Let’s say, for example, we’d like to see the top 10 customers by total sales and the relevant vice president at each client company. Let’s assume we have this data stored in a single table, structured like the records shown in Table 1.1. The records in this table might exist in any database as exact duplicates, but the way in which the database is asked for records can change drastically from system to system.

<table>
<thead>
<tr>
<th>VP First Name</th>
<th>VP Last Name</th>
<th>Customer</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>Doe</td>
<td>ABC Co.</td>
<td>1,000</td>
</tr>
<tr>
<td>Jane</td>
<td>Doe</td>
<td>XYC Inc.</td>
<td>500</td>
</tr>
</tbody>
</table>

Table 1.1 Customers with Sales Information
Now, let's look at some different SQL syntaxes, depending on the database that stores this table. For a table in Oracle, a developer would need to write a query that looks something like Listing 1.1. Oracle utilizes a double pipe (||) to concatenate strings and includes a useful rownum reserve name for tracking result set values, which can then be used.

```sql
Select VP_FIRST_NAME || ' ' || VP_LAST_NAME as VP_NAME,
       sum(sales) from table1 where rownum <= 10
       group by CUSTOMER order by sum(sales)
```

Listing 1.1 Oracle Syntax

For a table in Microsoft SQL Server, a developer would need to write a query that looks something like Listing 1.2. Microsoft SQL Server doesn’t have a rownum object that can be referenced; instead, the keyword top will select the top n number of records. Microsoft SQL Server also uses plus signs (+) for concatenation.

```sql
Select top 10
       VP_FIRST_name + ' ' + VP_LAST_NAME as VP
       Customer,
       sum(sales) from table1
       group by CUSTOMER order by sum(sales)
```

Listing 1.2 Microsoft SQL Server Syntax

For a table in PostgreSQL, you would write a query like the one in Listing 1.3. PostgreSQL, like Oracle, uses double pipes to tie strings together; however, unlike both Oracle and the Microsoft SQL Server, you’ll use a different keyword, limit, to restrict our result set to the top 10.

```sql
Select VP_FIRST_name || ' ' || VP_LAST_NAME
       Customer,
       sum(sales) from table1
       group by CUSTOMER order by sum(sales)
       limit 10
```

Listing 1.3 PostgreSQL Syntax

Even within the same database brand, differences among versions can also result in syntactical changes and, over time, through new releases, result in better ways to execute code. SAP Data Services enables ETL developers to ignore these differences in code, often without having to write any code at all.

The SAP Data Services user interface is primarily drag-and-drop. Rather than writing SELECT statements, although the option is available, you can import the table metadata and map columns from the source table to the target table by dragging and dropping columns and dragging. Queries are no longer lines of code but boxes that house all the individual configuration panels, dropdown menus, and function calls that make up a query. Once the configuration is satisfactory, the SAP Data Services application server executes the code by translating the configuration into the necessary SQL syntax required by both the source and target databases. An example of an SAP Data Services job is shown in Figure 1.1.

Figure 1.1 An Example SAP Data Services Job

For example, a common data transformation involves the location of a substring within a string. In SAP Data Services, similar to other programing languages, this transformation is known as an Index() function. Let’s say we have, as shown in Table 1.2, an example dataset that includes product codes and descriptions that no longer meet the business definition; thus, data manipulation is required.

<table>
<thead>
<tr>
<th>PRODUCT_CODE_LONG</th>
<th>PRODUCT_NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB-123</td>
<td>Cotton Swabs 500 Ct</td>
</tr>
<tr>
<td>KP-345</td>
<td>Cotton Swabs 1000 Ct</td>
</tr>
</tbody>
</table>

Table 1.2 Example Dataset
Perhaps a business requirement is to remove the text before the dash in a product code before sending the data to another system. A common solution for this in SAP Data Services is to leverage the index function along with a left trim (ltrim). The SAP Data Services code would look as follows:

```sql
ltrim(PRODUCT_CODE_LONG, 1, index(PRODUCT_CODE_LONG, '-', 1))
```

Regardless of the source database, this line of code will not require alternate syntax. With SAP Data Services, you don’t need to know that Oracle equivalent `Index()` function is called `Instr()` or that, to trim off the left side of a string in Microsoft SQL Server, the function `Right()` is required. Let’s not forget that this data might not be in a database at all! Instead, the data could be in an Excel file or even stored within a third-party cloud solution such as Salesforce.com. Regardless, SAP Data Services will determine the proper syntax required for the transformation logic.

If your organization needs to cast a wide net to unify numerous databases and perform complex data transformations, SAP Data Services is likely to be the preferred option. But what if your scope isn’t that wide? Other ETL tools are available to you, including one already built into the SAP HANA platform itself: SDI. However, to work with data not already inside SAP HANA, we’ll need to look at another component first, SAP HANA smart data access (SDA). While not specifically an ETL tool, we’ll discuss SDA because of its importance when leveraging SDI.

**SAP HANA Smart Data Access**

SDA is another piece of that SAP HANA platform. You might notice that this tool is not of specific to data provisioning. SDA provides a window into another database, thus allowing you to view and query without having to copy that data over to SAP HANA. The data never leaves its source system and is never written to the SAP HANA hard disk when leveraging SDA. However, you can see the data directly within your SAP HANA development environment under the **Provisioning** folder, as shown in Figure 1.2, which allows you to create remote sources and import virtual tables.

![Figure 1.2 SDA from the Provisioning Folder in SAP HANA Studio](image)

However, as you can probably guess, SDA’s virtual tables can be leveraged by SDI as source tables to facilitate an SAP HANA-based ETL solution, with, of course, some limitations. At the time of this writing, SDA in SAP HANA 2.0 includes the following 17 ODBC connections out of the box:

- ASE
- TERADATA
- IQ
- SAP HANA
- HADOOP
- GENERIC ODBC
- ORACLE
- MSSQL
- NETEZZA

![Figure 1.3 SDA Virtual Tables](image)
SDA also includes four destinations so you can leverage external procedure calls on your data when SAP HANA is not appropriate, for example, when using the open-source machine learning library TensorFlow or an rServe server. The four destinations are as follows.

- HADOOP
- SPARK SQL
- rSERVE
- GRPC

As long as these built-in ODBC connections meet your requirements, SDA might be all you need. SDI can simply refer to the virtual tables exposed by these SDA adapters as source tables, execute the SQL required, and then write the results to disk in SAP HANA. But, if you have source systems not accessible via the adapters listed above, one additional piece of software can be leveraged to extend beyond SDA’s predelivered ODBC adapters—SDI.

**SAP HANA Smart Data Integration**

Also an ETL tool, SDI offers much the same core functionality as SAP Data Services. SDI can leverage all the ODBC connections mentioned previously plus an additional 20 Java adapters have been developed by SAP and are distributed via the Data Provisioning Agent. Additionally, if these prebuilt solutions still don’t meet your needs, you can extend SDI’s integration further by writing your own Java adapter utilizing the SAP HANA Adapter software development kit (SDK).

One key difference between SDI and SAP Data Services is that, if you already have SAP HANA, you already have SDI. As a core component of the SAP HANA platform, every version of SAP HANA from SP 09 on has SDI built in and ready to deploy. If additional adapters are required, for example, for reading from a flat file or for connecting to a web service, you’ll need to complete an extra step first: You’ll need to deploy the Data Provisioning Agent, shown in Figure 1.4. The SAP HANA Data Provisioning Agent Configuration screen allows you to deploy 20 additional Java adapters to supplement the adapters already provided by SDA.

**Why a separate piece of software?** For SAP, this segregation of duties isolates the database from the data transfer mechanism and ensures that the processing power required by and promised to the SAP HANA system remains unaffected. Thus, SAP recommends utilizing a second server or a virtual machine (either Linux or Windows) to run the Data Provisioning Agent, from which your Data Provisioning Agent adapters will be deployed. Luckily, this free and lightweight piece of software can even be run locally on a typical developer’s laptop for testing purposes.

Another significant difference between the two tools is that, with the changes that have come with SAP HANA extended application services, advanced model (SAP HANA XSA) in SAP HANA 2.0, SDI development can be done completely in a web browser via the SAP Web IDE, as shown in Figure 1.5, which shows two tables being joined, but no output has been created. This web-based feature can greatly simplify processes and reduce the effort required for developer onboarding. Simply grant developers the appropriate role while creating their user and provide the link. No need to install client tools with the appropriate version, or even SAP HANA Studio or
Eclipse, the original development IDEs for SDI. SDI flowgraphs can be built using the SAP Web IDE, an SAP HANA XSA application accessible via a web browser.

Finally, the largest difference between the two tools involves their overall purposes. SDI’s purpose is to provision SAP HANA. Though packed with data federation options and extensibility via the SDK, SDA’s primary function is to load data into SAP HANA, not into other systems. While loading data into SAP HANA is probably your immediate goal, keep in mind your organization’s long-term goals. If loading an array of multiple databases other than SAP HANA is not a concern at the moment, SDI might be the perfect fit.

SDI is a feature-rich ETL solution capable of meeting many, if not all, of your SAP HANA provisioning requirements. In Chapter 2, we’ll cover how to get started developing SDI flowgraphs, how to set up the Data Provisioning Agent (as well as deploying its most common adapters), and how to leverage them in an SDI-based ETL solution. But, what if the data to be pulled into your SAP HANA environment isn’t quite up to par? As an aside, this book will also cover a few specific transformations within SDI in depth that call under their own acronym: SAP HANA smart data quality (SDQ).

### 1.1.2 Cleansing

While similar to ETL (and in the case of SAP Data Services bundled with cleansing tools), cleansing requires a different type of logic, something smarter. Where ETL tools will leverage joins by matching two keys exactly, cleansing leverages fuzzy joins and looks for likely matches with some degree of confidence. The goal of a cleansing tool is to find out whether a given piece of data captures the intent of the user who entered it. If you’ve ever been unlucky enough to have to join two datasets by something as fluid as company names (or worse, address lines), then you’ve experienced the challenges that come with programmatic cleansing. Take, for example, the records shown in Table 1.3. The number of ways different users might input the same address are staggering, and to a database, these variations are all equal in validity.

<table>
<thead>
<tr>
<th>Source System Name</th>
<th>Address Line</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloud CRM</td>
<td>293 1st Avenue</td>
</tr>
<tr>
<td>On Prem ERP</td>
<td>293 First Ave.</td>
</tr>
</tbody>
</table>

Table 1.3 Possible Data Inputs

To an analyst, these two addresses are clearly the same, but not so to a database. To avoid having to sift through millions of records, hunting for duplicates and valid links, you can leverage one of the tools in this category to ensure you’re making efficient use of your limited SAP HANA storage:

- SAP HANA smart data quality (SDQ)
- SAP Agile Data Preparation
- SAP Data Quality Management, microservices for location data

### SAP HANA Smart Data Quality

As a component of SDI, SDQ can be utilized to cleanse data already stored in SAP HANA, either in batch jobs during extractions from other systems or in real time as data becomes available to the SAP HANA system. SDQ is ultimately a subset of functions available to the SDI developer that can be included in flowgraphs, which is similar to the data quality transforms found in SAP Data Services, but only available with the appropriate license. While not as diverse as the data quality capabilities in SAP Data Services, SDQ is well suited for parsing and standardizing free-form text, without the need for an additional server, application, or licensing. However, you’ll need to take into account additional costs when cleansing address data is required. An annual subscription fee is required to access the most up-to-date address information across all SAP address cleansing solutions, including SAP Data Services. These address information files referred to as directories and are required for the different address cleansing engines to perform their logic. Once purchased, simply add the directories to the correct server location to enable validating and improving address data coming into your SAP HANA system.

Though only a subset of SDI, due to the numerous configurations required, we’ll explore SDQ extensively to ensure you get the most out of your decision to utilize
SDI as your SAP HANA provisioning tool. However, SDQ is not the only method that an organization can use to enhance data quality in their SAP HANA systems.

**SAP Agile Data Preparation**

SAP Agile Data Preparation, shown in Figure 1.6, is the most business analyst-friendly provisioning method discussed in this book. If you’re familiar with the self-service business intelligence trend popularized by tools such as SAP BusinessObjects Web Intelligence and SAP Lumira, SAP Agile Data Preparation extends the reach of that trend deeper into backend systems by offering business users an easy-to-understand web interface to connect data sources, whether a remote database or a local file, and perform common database tasks such as joins, formulas, and even cleansing. SAP Agile Data Preparation is, like SDI, an SAP HANA XSA application accessible with a web browser.

SAP Agile Data Preparation itself is ultimately an SAP HANA XSA application that, similar to SAP Data Services, translates a user’s configurations, transformation, and cleansing rules into backend SQL commands. However, these commands are not limited by user sessions in any way. Rather than obscuring a user’s “development” behind the finished product, the process itself is exportable. Once a user has written code, this code can be saved and shared to improve reusability and standardization. Exporting an SAP Agile Data Preparation job shows the underlying commands generated, which are in fact SDI flowgraphs. Thus, these flowgraphs can be sent to IT as a prototype, enabling IT to better understand what the business needs really are and to improve the development process.

SAP Agile Data Preparation, while an extension of the SAP HANA platform, does not however actually require an SAP HANA instance. SAP also offers an SaaS SAP Agile Data Preparation solution via the SAP Cloud Platform. We’ll cover how to set up both on-premise and cloud SAP Agile Data Preparation in depth in Chapter 4.

**SAP Data Quality Management, Microservices for Location Data**

In addition to SAP Agile Data Preparation, SDQ and the data quality transforms found in SAP Data Services, we’ll be covering one final data quality product, SAP Data Quality Management, microservices for location data. Microservices are much like they sound, micro. Microservices are application programming interface (API) endpoints that do one thing and one thing only. This granularity allows developers plug in services as needed and allows the owners of the service to easily manage and debug them. SAP announced its foray into the microservices realm by pulling out the most complicated pieces of the ETL process, address cleansing and geocoding.

Through a cloud service, you can visit the microservices web page to view usage, billing, and connection information (see Figure 1.7). However, in order to actually leverage the service, you need to integrate programmatically through SAP Data Services or another application backend.

Figure 1.6  SAP Agile Data Preparation User Interface

Figure 1.7  SAP Cloud Platform Cockpit Microservices Page
As we’ll see in Chapter 3 and Chapter 5, these processes offer numerous options and require annual updates. If these setup costs, both in time and money, seem prohibitive, the microservices route might be a better choice instead. We’ll walk you through the simple process of setting up your microservices account, as well as some common use cases, and describe the integration process of using SAP Data Quality Management microservices into common applications.

1.1.3 Replication

The final category of data provisioning tools is also the simplest. Replication is the purest form of data transference: Table A in System 1 should match Table A in System 2. Complexity comes into play during execution. How often is System1/TableA updated? How often should System2/TableA be refreshed? Should System1 push the data to System2, or should System2 pull the data from System1? How will you detect changes in System1? These questions can be answered by a replication tool. Not included in the following is SAP Data Services, where replication via real-time jobs can be achieved, but these other tools require much less development to implement:

- SAP Landscape Transformation Replication Server
- SAP HANA smart data integration (SDI)

With this grouping mind, you should have a clear understanding of where to direct your attention given a particular use case and the tools available to you. Use Table 1.4 to quickly determine the right tools, based on the type of provisioning and business need, for either batch (B) (i.e., periodic) processing or real-time (RT) (i.e., immediate) processing. Please note that SDQ is a component of SDI; thus, technically, SDI performs cleansing functions as well.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Manipulate</th>
<th>Copy</th>
<th>Cleanse</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAP Data Services</td>
<td>B/RT</td>
<td>B</td>
<td>B/RT</td>
</tr>
<tr>
<td>SAP HANA smart data integration (SDI)</td>
<td>B/RT</td>
<td>B/RT</td>
<td>B/RT</td>
</tr>
<tr>
<td>SAP HANA smart data quality</td>
<td>B/RT</td>
<td></td>
<td>B/RT</td>
</tr>
<tr>
<td>SAP Data Quality Management</td>
<td>B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAP Agile Data Preparation</td>
<td>B</td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>SAP Landscape Transformation Replication Server</td>
<td>RT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.4 Tools for Batch and Real-Time Capabilities

SAP Landscape Transformation runs on the SAP NetWeaver stack. Trigger-based replication has been a staple of many database architectures for years; however, just like SQL has its own flavors, replication too can vary by database brand and version, in this case SAP ERP and SAP Business Warehouse, on which your application is installed. The SAP LT Replication Server fills the gap nicely at the application level, much like SAP Data Services, but with a core focus on real-time replication rather than ETL.

SAP LT Replication Server provides a cockpit view for setting up tables to be initialized, replicated, and reloaded. Generally, once set up, you shouldn’t need to revisit the cockpit outside of occasional maintenance or troubleshooting, as shown in Figure 1.8.

Figure 1.8 SAP LT Replication Server Cockpit View

True, for some transformation capabilities, all of which we’ll cover in this book, the SAP LT Replication Server shines in its ability to simplify the replication of SAP data into a target enterprise data warehouse (EDW). In this chapter, we’ll dive into what capabilities exists, how we can leverage these capabilities to generate real-time views of our data, and when best to leverage the SAP LT Replication Server in your provisioning strategy.

1.2 How Are These Tools Used Together?

Now that we’ve touched on each tool individually, you should understand why using all of these tools to their fullest extent within a single organization is rather unlikely. In fact, with so many overlapping functionalities, more likely, only two or three of
these tools will be heavily utilized in a production scenario. While we’re used to seeing some common pairings, ultimately every environment will require a different combination tool.

One of the most challenging decisions for anyone new to the SAP EIM space is deciphering when to utilize one or more of the ETL tools described in this book. While these tools overlap in many ways, each of them excel in one or more areas that the others aren’t designed to support. Over the years, the authors have come to rely upon the following three criteria in order to arrive at the appropriate mix for a given environment:

- **Scope:** How many unique data storage solutions are within the scope of your provisioning strategy?
- **Quality:** How much transformation, cleansing, and manipulation is required before the data becomes meaningful/useful?
- **Latency:** How quickly must the target system (SAP HANA in the case of this book) be updated relative to the data being written to the source system?

Simply asking these three questions often requires booking a conference room for a week. As depicted in Figure 1.9, none of these questions are meant to build on the other, and not all of them will hold equal weight in the final tool mix your organization decides on.

![Figure 1.9 Latency, Quality, and Scope](image)

The following three matrices, Table 1.5, Table 1.6, and Table 1.7, can help you narrow down the optimal tool mix for your situation.
**1. Introduction**

For example, let’s assume that after reviewing the requirements our scope, quality, and latency, we determine that we wish to utilize SAP HANA as our EDW, with no separate staging or archival system. We acknowledge that, after reviewing the sources of our data, some manipulation will be required to unify the systems, but not much, and that our users are comfortable with nightly data refreshes. As a result, we see SDI and SAP Data Services support all three requirements, with SAP Data Services offering more capability when it comes to data quality and manipulation. If we are not confident in our quality assessment, we might lean more towards SAP Data Services, however, in this scenario we are at least certain that neither SAP LT Replication Server nor SAP Agile Data Preparation will meet our needs.

That said, by far the most common scenario we’ve seen is leveraging SAP LT Replication Server and SAP Data Services to provide near real-time reporting outside of SAP ERP. This scenario is probably prevalent because of the popularity of the SAP HANA sidecar architecture, which enables SAP customers to query massive volumes of SAP ERP transactional data directly, without having to reinstall and migrate their SAP ERP environment. Instead, SAP LT Replication Server (or sometimes SAP Data Services batch jobs) can replicate the data to SAP HANA tables.

However, often, customers still need to use “helper tables,” tables that provide flags and other user information, to get the most out their transactional data. Thus, SAP Data Services provides batch processing to generate keys, perform lookups, and fill in other gaps that neither the SAP LT Replication Server nor SAP HANA views could effectively resolve.

### Table 1.7 Utilize this Table to Determine which Tools Best Support your Latency Requirements

<table>
<thead>
<tr>
<th>Capability</th>
<th>Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SAP Data Services</td>
</tr>
<tr>
<td>Batch Processing</td>
<td>Great</td>
</tr>
<tr>
<td>Real-Time</td>
<td>Great</td>
</tr>
<tr>
<td>Processing</td>
<td>Real-Time Replication</td>
</tr>
</tbody>
</table>

Of course, nothing prevents you from leveraging SDI to do the same thing as SAP Data Services in some scenarios. Further, of course, due to its integration capabilities, if you’re using SAP Agile Data Preparation, you’ll probably want to leverage the export process to flowgraph functionality for developing reusable and standardized logic. Ultimately, the architect is the one to decide, while system administrators and business users must decipher which tools should be utilized for which purposes.

**Example**

Let’s look at a hypothetical use case where every tool plays a role within an imaginary enterprise information management team at a large international organization, MaxWidgets, Inc.

MaxWidgets is a large organization that has grown via several international acquisitions. As a result, numerous ERP and EDW systems are spread throughout the world, the largest of which are in Beijing, Ireland, and Memphis, TN. The executive team is struggling to get a clear picture of total sales by region because each region has their own method of collecting sales data. Some data is easy and comes in via the online store, but many customers visit local branches and make purchases through in-person sales representatives, who, unfortunately, aren’t patient with the CRM tool. The deliveries, especially in Beijing, are often managed by individual reps and rarely tie back to the billing address on the order. While the Memphis and Ireland sales data is pretty consistent, these branches have far more sales and generate several times the amount of records per day, compared to the Beijing branch.

Now, let’s say that leadership has decided to move all sales data into SAP HANA; however, not all of the data is created equal. We already know the address data in Beijing has tons of duplicates and errors as the sales reps key in only the bare minimum into the CRM to complete the opportunity entry, but Memphis is running a leg-acy SAP ERP system on old hardware, and Ireland has a homegrown BI application that only publishes on-demand reports that are essentially stored procedures that call back to JavaServer Pages (JSPs).

Digging into the Ireland BI application, you realize that a massive ETL effort is required to recreate the stored procedure and JSP logic. You decide to put all of your SAP Data Services resources on the task, and slowly but surely, you begin extracting the Ireland data straight into your SAP HANA tables. However, you can’t afford to wait on an available SAP Data Services resource to begin work on the Beijing and Memphis data, so you turn to your SAP HANA team for assistance. They propose pulling the Beijing data via SDI; however, they recommend cleaning up the data in transit. Not much more transformation is required outside of the cleansing, and you...
don't own Beijing address directories, so you decide to keep the SDI layer simple for now and instead use the SAP Data Quality Management microservice for Beijing. In this way, if you decide to convert the Beijing sales data to an SAP Data Services job, switching over will be easier.

With Beijing and Ireland out of the way, you turn your sights to the legacy SAP system in Memphis. They've been talking about upgrading the system for years, but haven't gotten around to it. You know what tables you need, but nightly batches would strain the old servers, so you decide to leverage SAP LT Replication Server and replicate each record as it comes in in real time. SAP Basis gets you up and running, but then you realize something is off about the customer master—it seems old. Turns out the business has been maintaining the customer master outside of SAP through a combination of Excel files and Microsoft SQL Server databases that reference SAP document numbers. After all, the old system has been "about to go away" for years. Rather than trying to piece these files together with the few SAP Data Services developers you have available, you decide to use SAP Agile Data Preparation and allow the business to continue to map sales headers to their SQL database. This slight change to their current process still should reduce the number of Excel files floating around, and that's something everyone can get on board with.

1.3 Summary

In this chapter, we focused on the high-level strengths of each tool, providing a pretty thorough inventory of the provisioning options available for SAP HANA from SAP. In the next few chapters, we'll take a close look at each of these applications, describe how to get started working with them, and discuss some common pitfalls you may encounter along the way. First, let's focus on SDI, including how to get it up and running and how to get started provisioning SAP HANA.
# Contents

Preface ........................................................................................................................ 13

1 Introduction .................................................................................................................. 17

1.1 What Are the Tools for Provisioning Data? ............................................................. 17
    1.1.1 Extract, Transform, and Load ........................................................................... 18
    1.1.2 Cleansing ................................................................................................................ 26
    1.1.3 Replication .............................................................................................................. 30

1.2 How Are These Tools Used Together? ........................................................................ 31

1.3 Summary ..................................................................................................................... 36

2 SAP HANA Smart Data Integration ........................................................................... 37

2.1 What Is SAP HANA Smart Data Integration? .......................................................... 37

2.2 Use Cases for SAP HANA Smart Data Integration ...................................................... 38

2.3 Installation and Configuration ................................................................................... 39
    2.3.1 Data Provisioning Server .................................................................................... 40
    2.3.2 Data Provisioning Delivery Unit ..................................................................... 41
    2.3.3 Data Provisioning Agent .................................................................................... 44

2.4 Using SAP HANA Smart Data Integration ................................................................. 48
    2.4.1 SAP HANA Web-Based Development Workbench .................................... 48
    2.4.2 Creating Flowgraphs ........................................................................................... 50
    2.4.3 Configuring the Data Provisioning Agent for Flat File Access ............... 54
    2.4.4 Reading Flat Files ............................................................................................... 57
    2.4.5 Building Blocks ................................................................................................... 67
    2.4.6 Real-Time Flowgraphs ....................................................................................... 78
    2.4.7 Monitoring ........................................................................................................... 83

2.5 Summary ..................................................................................................................... 89
## 3 SAP HANA Smart Data Quality

### 3.1 What Is SAP HANA Smart Data Quality? ....................................................... 91
### 3.2 How Do SAP HANA Smart Data Integration and SAP HANA Smart Data Quality Work Together? .......................................................... 92
### 3.3 Installation and Configuration ................................................................... 93
  3.3.1 Enabling the Script Server ................................................................. 93
  3.3.2 Downloading and Deploying SAP Smart Data Quality Directories .... 95
  3.3.3 Creating Authorized Users for SAP Smart Data Quality .................. 101
### 3.4 Using SAP HANA Smart Data Quality ...................................................... 103
  3.4.1 Identifying Cleansing Options ......................................................... 103
  3.4.2 Identifying Matching Options ......................................................... 110
  3.4.3 Identifying Geocode Solution Options .......................................... 117
  3.4.4 The Script Server ............................................................................. 121
### 3.5 Summary ................................................................................................... 122

## 4 SAP Agile Data Preparation

### 4.1 What Is SAP Agile Data Preparation? ...................................................... 123
### 4.2 SAP Agile Data Preparation and SAP HANA .......................................... 124
### 4.3 SAP Agile Data Preparation: On-Premise versus Cloud ....................... 124
### 4.4 Installation and Configuration .................................................................. 126
  4.4.1 Downloading the Files ....................................................................... 126
  4.4.2 Importing the Delivery Units ............................................................. 132
  4.4.3 Adding Data Domain Tiles ................................................................. 138
  4.4.4 Security Management ....................................................................... 139
### 4.5 Using SAP Agile Data Preparation ............................................................ 140
  4.5.1 Creating a Project and Loading Data ................................................. 140
  4.5.2 Navigating the Side Panel ................................................................. 145
  4.5.3 Reviewing Data Quality Statistics .................................................... 147
  4.5.4 Actioning Data .................................................................................. 149
  4.5.5 Cleansing and De-duplicating Data .................................................. 156

## 5 SAP Data Services

### 5.1 What Is SAP Data Services? ..................................................................... 167
### 5.2 Installation and Configuration .................................................................. 168
  5.2.1 Install Information Platform Services .............................................. 172
  5.2.2 Install SAP Data Services ................................................................. 194
### 5.3 Using SAP Data Services ......................................................................... 202
  5.3.1 Batch Data Loading .......................................................................... 202
  5.3.2 Best Practices .................................................................................... 211
### 5.4 Summary .................................................................................................. 217

## 6 SAP Landscape Transformation Replication Server

### 6.1 What Is the SAP Landscape Transformation Replication Server? .......... 219
### 6.2 Installation and Configuration .................................................................. 222
  6.2.1 ABAP Source System ...................................................................... 223
  6.2.2 Separate Server with an ABAP Source System ................................ 224
  6.2.3 Separate Server with a Non-ABAP Source System .............................. 224
### 6.3 Using the SAP LT Replication Server ....................................................... 225
  6.3.1 Configuring and Managing the Replication Process ......................... 230
  6.3.2 Creating a Configuration ................................................................... 232
6.3.4 Initial versus Ongoing Data Replication ....................................................... 234
6.3.5 Transformation Capabilities ....................................................................... 236

6.4 Summary .................................................................................................... 238

7 SAP Data Quality Management, Microservices for Location Data ................. 241

7.1 What Is SAP Data Quality Management, Microservices for Location Data? ........ 241
7.2 Invoking Microservices for Location Data .................................................. 243
  7.2.1 Address Cleansing and Geocoding ......................................................... 243
  7.2.2 Reverse Geocoding .............................................................................. 249
  7.2.3 Information Codes and Messages ......................................................... 251
7.3 Installation and Configuration .................................................................... 252
  7.3.1 Getting Started ..................................................................................... 252
  7.3.2 Supported Integrations ....................................................................... 253
  7.3.3 Authentication .................................................................................... 256
  7.3.4 Configuration Editor .......................................................................... 257
7.4 Using Prebuilt Functions ............................................................................. 258
7.5 Summary .................................................................................................. 259

8 SAP HANA Data in the Cloud ......................................................................... 261

8.1 Cloud Considerations .................................................................................. 261
8.2 SAP Cloud Platform .................................................................................... 265
  8.2.1 SAP Cloud Connector ........................................................................ 265
  8.2.2 Architecture ....................................................................................... 267
  8.2.3 Integration ......................................................................................... 268
8.3 Amazon Web Services ............................................................................... 270
8.4 Microsoft Azure ......................................................................................... 275
8.5 Summary .................................................................................................. 279

9 Data Provisioning Case Studies ..................................................................... 281

9.1 Data Preparation for an Omnichannel Initiative ............................................ 281
  9.1.1 Company Background .................................................................... 282
  9.1.2 Solution .......................................................................................... 284
9.2 Supply Chain Analytics for Reducing Cost of Goods Sold ......................... 303
  9.2.1 Company Background .................................................................... 304
  9.2.2 Solution .......................................................................................... 307
9.3 Profile and Transform Customer Data .......................................................... 323
  9.3.1 Company Background .................................................................... 323
  9.3.2 Solution .......................................................................................... 324
9.4 Cleaning and De-duplicating a Mailing List ................................................. 332
  9.4.1 Company Background .................................................................... 332
  9.4.2 Solution .......................................................................................... 333
9.5 Summary .................................................................................................. 343

The Authors ................................................................................................. 345
Index ............................................................................................................ 347
Index

_SYS_REPO ........................................ 66-67, 81, 102

A

ABAP source system ......................................... 233
Access plans ...................................................... 236
Adapters ............................................................. 47, 57, 60, 83, 315
Address cleansing .............................................. 243
Address directories ............................................... 93
Address formats ................................................. 245
Address validation ............................................. 258
Addresses ................................................................ 27, 161
AFL ................................................................. 78
Agent Monitor ..................................................... 43, 83–84
Agents ................................................................. 46, 83
Aggregating data ................................................... 152
Aggregation nodes .................................................. 72–74
Amazon Web Services (AWS) ......................... 125, 270–271
vs Microsoft Azure .............................................. 276
API Management Console .................................. 268–269
API requests .......................................................... 243
request properties ............................................... 244
response properties ........................................... 247
Application Designer .......................................... 265
Application function libraries ............................ 121
Application function modeler ............................. 91
Application programming interface (API) ....... 29
Association Editor ............................................... 301
Associative match ............................................. 299–303
Attribute change package .................................. 254
Authentication .................................................... 256
client certificate .................................................. 254
Authorizations ................................................... 232, 234

B

Batch ................................................................. 30, 34, 78, 81, 87, 202
Batch data loading ............................................. 202, 211
Batch jobs ........................................................... 172, 193, 204
Bill of material (BOM) ......................................... 506
Blueprint packages ............................................ 256
Break group key ................................................... 296, 299
Business configuration sets .................................. 254
Business intelligence (BI) ..................................... 35

C

Calculation views .............................................. 165
Case studies ....................................................... 281
customer data ..................................................... 323
mailing list ........................................................ 332
omnichannel retail ............................................. 282
supply chain analytics ......................................... 303
Case transforms ................................................. 188, 216
configuration ...................................................... 189
Catalog .................................................................. 23, 38, 49, 61
Central Management Console (CMC) .............. 142
Central Management Server (CMS) ................ 198
Change data capture (CDC) .............................. 78, 81, 205
Checkpoint recovery ........................................... 176–177
Cleanse transform ............................................. 93, 95, 103, 105, 110, 117
Cleansing ......................................................... 26–28, 156, 160, 282, 285–286,
dictionaries ....................................................... 161
options ............................................................... 103
Clients .................................................................. 310
Cloud ................................................................. 19, 29, 46
Cloud deployments ............................................ 262
Cloud migration ................................................... 263
Cloud providers ................................................... 262
Cluster tables ...................................................... 235
Configuration and Monitoring Dashboard .......... 226
Configuration Editor .......................................... 252, 257
Consolidated customer ........................................ 284
Consumption-based pricing model ..................... 242
Containerization ................................................... 263
Content Management Server (CMS) .............. 124
Credentials mode ............................................... 58
cron ..................................................................... 86
CSV ..................................................................... 55, 165, 333
Customer relationship management (CRM) .... 35

D

daemon.ini ......................................................... 40, 94
Data cleanse ....................................................... 92
Data compression .............................................. 207
Data enrichment ............................................... 146

347
Index

Data federation .............................. 23, 37
Data flows ........................................ 183–184, 188, 212
Data Integrator .................................. 19
Data manipulation ........................... 149
Data mart ..................................... 208, 210
Data Migration Server (DMIS) add-on .... 222
Data modeling .................................. 122
Data provisioning ............................ 312
Data Provisioning Agent 18, 24–25, 40–41, 43–44, 47, 54–59, 81, 83–84, 125, 127
Data provisioning server ................... 40
address cleansing ..................... 243
assessment .............................. 148
genealogy .................................. 248
reversion-genealogy .................... 249
statistics .................................... 147
Data sink ...................................... 79, 121
Data Source Browser ................. 143
Data sources ................................ 113, 142
Data structures ............................. 209
Data warehouse ............................ 215, 308, 318, 321
Database connection ...................... 221
Database management system (DBMS) .... 263
Database triggers ......................... 219
Dataflows ................................... 288, 292–293
Datastores .............................. 168, 255–256, 287
configuration properties .......... 169
connection parameters .............. 168
eample .................................... 170
Date generation .......................... 78
DB2 system .................................. 314
De-duplicating .......................... 156–157, 342–343
Delivery units ...................... 40–41, 43, 125, 136
import ....................................... 132
installer .................................. 135
Dimension .................................. 77
Dimension tables .......................... 179
Direct Connect ............................ 271
Download Manager ...................... 45
Dq_reference_data_path .............. 100

E

Eclipse ......................................... 26
Editor ....................................... 52, 60–61
Elastic Compute Cloud (EC2) ............ 270
Endpoint ..................................... 265
Enterprise data warehouse (EDW) ....... 31, 35, 62, 332–333
Enterprise information management portfolio ........................................ 91, 242
Enterprise Semantic Services .............. 127
ETL .................................. 17–20, 23–24, 26, 29, 35, 37–38, 50, 67, 82, 89, 91, 122, 208
business rule enforcement stage .......... 216
driver stage .................................. 212
lookup stage .................................. 214
processing stage ........................... 213
Event-based rules ......................... 238
Excel .......................................... 36
Expression Editor ......................... 71
F

Fact tables .................................. 179
Field validations .......................... 192
Field-based rules ......................... 238
File adapter ................................ 54, 56, 58–60, 334
Filter transform .......................... 93
Filters ....................................... 67–68, 70–72, 74, 77, 341–342
node ........................................... 79
Flat files ..................................... 57, 60, 314, 325, 343
Dictionaries ............................... 154
FTP ........................................... 19
Fully qualified domain name .......... 231
Fuzzy joins .................................. 26
Fuzzy logic .................................... 110
Fuzzy match .................................. 159
G

Geocode ...................................... 244, 321–323
Geocode transform ..................... 95, 103, 117, 119–120
Git ........................................... 55
GUID ........................................ 232

H

Hadoop ...................................... 24
Harmonize values .......................... 151
hdflflowgraph ............................. 52
hdbserver ................................... 40
HDFS ........................................ 58
Hybrid solution ........................... 263

I

Import ........................................ 43
Index server .................................. 37
Information codes and messages .......... 251
Information Platform Services (IPS) ... 194, 197
Information Platform Services server .... 263
Initial Load .................................. 234
Input type .................................... 79
IT landscape .................................. 91

J

Java ............................................. 24
JIT Data Preview ....................... 68, 335, 340–341
Job server engine .......................... 215
Jobs ........................................... 172, 178, 293
Joints ......................................... 75–78, 300, 322, 330
node ........................................... 79, 77
JSON ........................................... 244

K

Kerberos .................................... 59

L

Latency ...................................... 220
Launchpad ........................................ 44
Linux .......................................... 25, 45
Logging tables .............................. 326, 328–329
Lookups ...................................... 215, 331–332
Ltrim (left trim) ............................ 22

M

Mapping .................................... 120, 289, 291, 327–328
Mass transfer .................................. 308
Match policy ................................... 115, 157
Match rule ....................................... 114
Match settings ................................ 115
Match transform ........................... 110, 112, 114
Multidatabase container (MDC) .......... 40

N

Neteza ......................................... 287
Nodes .......................................... 67–68, 72
Notifications .................................. 87–88

O

OAuth client .................................. 256
OData .......................................... 268
ODBC ........................................ 19, 23–24, 41, 57, 315
OLTP ........................................... 308
Ongoing replication ........................ 236
On-premise ...................................... 19, 29, 261
Oracle ........................................... 20, 314
Output types .................................. 79

P

Parallel workloads .......................... 175
Performance options ...................... 238
Personal security environment .......... 254
Pivot ........................................... 78
Schemas 59, 61, 66, 81, 288, 310, 313–314, 318, 320, 323
Script server ............................................... 93–94, 121
SDQ_USER .......................................................... 102
Security ................................................ 49, 55, 81, 139, 305
role ......................................................... 49
Sender queue ............................................... 235
Series execution ........................................ 178
Server Intelligence Agent (SIA) ...................... 198
SFTP ...................................................... 19, 130
Sharing data ............................................... 163
Sidebar .......................................................... 34
Single-use script object .................................. 182
SMTP .......................................................... 88
SOAP .......................................................... 41, 57
Social media ............................................... 282
Software development kit (SDK) .................... 24, 26
Software-as-a-service (SaaS) ......................... 29, 254
SQL ...................................................... 19–20, 31, 40–41, 60–61, 68–69, 74, 81–82, 85, 336
SQL Console ............................................. 104, 110, 130
SSO .......................................................... 59
Staging ...................................................... 75, 205, 207, 288, 334
Stateless application constructs ..................... 192
Storage .......................................................... 65
Suggestion lists .......................................... 247
Support Package Manager ................................ 254
Survival rules ............................................... 115, 158
SYSTEM user ............................................... 100
Table settings .............................................. 238
Tables ....................................................... 143, 227
Target tables ............................................... 121
Task Monitor ............................................... 43, 83–85
Technical user ............................................. 58
Template tables ........................................... 61–66, 109, 116, 119, 121
Tenants ..................................................... 40
Tensorflow ................................................... 24, 38
Traces .......................................................... 49
Transaction .................................................

LTR ......................................................... 226
LTBC ...................................................... 227, 235, 309
LTRO ...................................................... 228
LTRS ...................................................... 227

Transactional data ........................................ 34
Transformation capabilities ........................... 236
Trigger-based replication ................................ 219
Triggers ..................................................... 81
Truncate .................................................... 62
table .......................................................... 66
Try and catch block ...................................... 173

U
unpivot ...................................................... 78
Upsert ...................................................... 66
URL .......................................................... 46, 49–50
User roles ................................................... 226, 233

V
Validation transforms .................................... 190
classification ............................................. 191
Virtual private cloud (VPC) ......................... 270
Virtual private network (VPN) ...................... 262
Virtual tables ............................................ 23–24, 57, 59–62, 64, 78, 81, 318

W
Web service ............................................... 57
Weighted scoring ...................................... 293
WHERE clause ......................................... 185
Windows ................................................... 25, 45
Work process ............................................ 232
Workflows .................................................. 174, 177, 182, 288
failure ...................................................... 171
parallel execution ...................................... 175
series execution ........................................ 178
Worksheets .............................................. 310, 325, 327, 329, 331
Workspace ............................................... 53, 68, 72

X
XML web services ........................................ 173
Megan Cundiff is a data and analytics consultant at Protiviti where she works with clients from all industries to understand complex business challenges and implement end-to-end business intelligence solutions.

Vernon Gomes is a former IT industry systems administrator turned BI consultant. He is currently a senior consultant at Protiviti for data and analytics and is using his IT experience to assist clients in developing BI and cloud solutions.

Russell Lamb is a manager at Protiviti who has spent the last several years empowering organizations to use SAP HANA by enhancing their enterprise data warehouses, analyzing unwieldy SAP ERP tables, cleansing and storing SaaS-sourced CRM data, and extending their landscape into the cloud.

Don Loden is a managing director of data and analytics at Protiviti, with full lifecycle data warehouse and information governance experience in multiple industries. He is an SAP Certified Application Associate of SAP Data Services, and the author of three books and twelve articles on data management topics.

Vinay Suneja is a manager at Protiviti with more than five years of experience in implementing analytic solutions for clients in the retail, utilities, public sector, and banking industries. He is proficient with SAP BusinessObjects BI/SAW BW as well as big data technologies including SAP Lumira, SAP HANA, and Hadoop.

We hope you have enjoyed this reading sample. You may recommend or pass it on to others, but only in its entirety, including all pages. This reading sample and all its parts are protected by copyright law. All usage and exploitation rights are reserved by the author and the publisher.