Cloud Analytics on Google Cloud Platform with BigQuery and Informatica Intelligent Cloud Services

Reference Architecture
About Informatica

Digital transformation changes expectations: better service, faster delivery, with less cost. Businesses must transform to stay relevant and data holds the answers.

As the world’s leader in Enterprise Cloud Data Management, we’re prepared to help you intelligently lead—in any sector, category or niche. Informatica provides you with the foresight to become more agile, realize new growth opportunities or create new inventions. With 100% focus on everything data, we offer the versatility needed to succeed.

We invite you to explore all that Informatica has to offer—and unleash the power of data to drive your next intelligent disruption.
# Table of Contents

Abstract ................................................................................................................................. 4

Informatica Intelligent Cloud Services Overview .......................................................... 4

Google Cloud Platform Overview .................................................................................... 5
  Google Compute Engine .................................................................................................. 5
  Google Cloud Storage ................................................................................................. 5
  Google BigQuery .......................................................................................................... 5
  Google Cloud SQL ........................................................................................................ 6
  Google Virtual Private Cloud ...................................................................................... 6

The Case for Using Informatica Intelligent Cloud Services and Google BigQuery for Cloud Analytics ............................................................ 6
  Informatica Intelligent Cloud Services ....................................................................... 7
  Six-Step Synchronization Task Wizard ...................................................................... 7
  Mapping Designer ........................................................................................................ 8
  Secure Agent ................................................................................................................ 8
  Connectivity Options .................................................................................................... 9
  Google BigQuery .......................................................................................................... 9
  Google Cloud Platform ............................................................................................... 11
  Cloud Storage ............................................................................................................. 11
  Common Use Cases ..................................................................................................... 11

Deployment Steps ............................................................................................................ 13
  Informatica Intelligent Cloud Services Setup .......................................................... 13
  Secure Agent ................................................................................................................ 13
  Google Compute Engine and Networking .................................................................. 14
  Secure Agent Group ..................................................................................................... 14
  Data Source and BigQuery Connection Setup ............................................................ 15
  Google BigQuery Connection Setup .......................................................................... 15
  Google Cloud Storage Connection Setup ................................................................... 18
  Data Integration Task ................................................................................................... 19
  Mapping Designer ........................................................................................................ 28

Advanced Topics .............................................................................................................. 32
  BigQuery RECORD Data Type Support ....................................................................... 32
  Informatica PowerCenter Customers .......................................................................... 37

Summary .............................................................................................................................. 42
Abstract

Data-driven digital transformation initiatives leverage analytics technologies, and the power of trusted and timely data to deliver new insights, innovation, and faster decision making. The result is greater business agility and an enhanced competitive edge. New cloud-based technologies for analytics make it economically possible to ask questions involving a wide variety of data sources in ways that were never before possible.

For example, analyzing data collected about the sales of products and services makes it easier to spot meaningful insights about customer behaviors, and to predict such things as purchasing trends, next best steps, and usage patterns. Armed with deeper and more timely knowledge about what’s working and what isn’t, companies are better equipped to modify their products, services, and processes to fit their customers’ expectations and preferences. Understanding this kind of data better helps businesses disrupt their competitive landscape and develop new revenue streams by creating products and services from insights revealed by analyzing these patterns.

Three tools to help you achieve these goals are Informatica Intelligent Cloud Services™ (IICS), Google BigQuery, and Google Cloud Platform (BigQuery is part of the Google Cloud Platform). This reference architecture provides an overview of these tools and includes some common use cases, such as data warehouse modernization and big data analytics. This document also highlights best practices for applying IICS and Google BigQuery to the challenges of cloud analytics, such as accommodating a massive (and growing) volume of data from various sources, data cleansing and transformation, and managing continuously changing requirements.

This document is not a replacement for official statements of supportability. Rather, it is a guide to assist with deployment. Statements regarding configurations in the reference architecture are informational and should be cross-referenced with the most recent published documentation.

Informatica Intelligent Cloud Services Overview

Informatica Intelligent Cloud Services is a multi-tenant integration platform as a service (iPaaS) delivered as a subscription service that provides cloud data integration, process integration, API management, and prebuilt connectivity, among other data management capabilities. This modern “cloud of clouds” delivers industry-leading data management and integration capabilities, powered by artificial intelligence (AI). It enables you to develop and run data integration tasks or mappings, task flows, scheduling, monitoring, and notifications in an integrated, intuitive, next-generation browser interface. It is also modular, so you can start small and expand data management capabilities as your needs grow.

Informatica Intelligent Cloud Services includes the following components:

Informatica Intelligent Cloud Services

A hosted platform service that stores metadata related to all integration tasks, connections properties, and organization information. Connections properties, integration tasks, and mappings are configured through an intuitive visual designer in a web browser.
Informatica Intelligent Cloud Services Assets
Purpose-built integration assets such as Synchronization, Replication, PowerCenter® Tasks, Mass Ingestion Tasks, Mappings, and Taskflows. Templates for common integration patterns are also included.

Informatica Intelligent Cloud Services Secure Agent
A lightweight, self-upgrading software application that securely administers data integration tasks. Deployed inside your network or on Google Compute Engine, it expedites secure communications across the firewall between your organization and Informatica. Informatica Intelligent Cloud Services integration services use the Secure Agent to traverse the firewall and access sources and targets such as applications, relational databases, and file stores over the network.

Informatica Intelligent Cloud Services provides native, high-performance data and application integration with Google BigQuery and supports out-of-the-box integrations with hundreds of cloud and on-premises databases, data warehouses, and applications including Google Cloud Storage, Cloud SQL, and Cloud Spanner.

Google Cloud Platform Overview
Google Cloud Platform (GPC) consists of a set of physical assets, such as computers and hard disk drives, coupled with virtual resources, such as virtual machines (VMs), that are housed in Google’s data centers around the globe. Access to these physical and virtual resources on GCP is exposed through services such as IICS, which is compatible with the following GCP offerings:

Google Compute Engine
Compute Engine delivers virtual machines running in Google’s innovative data centers connected by a worldwide fiber network. Compute Engine supports scaling from single instances to a global, load-balanced, cloud computing infrastructure. Compute Engine supports Windows and Linux OS flavors, and comes in both preconfigured and custom configurations. Documentation.

Google Cloud Storage
Cloud Storage facilitates worldwide storage and retrieval of any amount of data at any time. Cloud Storage is suited for a range of scenarios, including storing content for data analytics, archival purposes, and disaster recovery, or distributing large data objects that users need to download directly. Documentation.

Google BigQuery
BigQuery is a powerful big data analytics platform used by all types of organizations, from startups to Fortune 500 companies. It provides a fully managed, petabyte-scale, low cost, and serverless enterprise data warehouse for analytics. With no infrastructure to manage and no database administrator required, it enables you to focus on analyzing data to find meaningful insights using a familiar SQL interface. Documentation.
Google Cloud SQL

Cloud SQL provides a database infrastructure for cloud applications that makes it easy to set up, maintain, manage, and administer relational PostgreSQL and MySQL databases in the cloud. Cloud SQL offers high performance, scalability, and convenience. [Documentation](#).

Google Virtual Private Cloud

Virtual Private Cloud (VPC) is a comprehensive set of Google-managed networking capabilities, including granular IP address range selection, routes, firewalls, VPNs, and Cloud Router. VPC lets you provision GCP resources, connect them to each other, or isolate them from one another. You also have options to define fine-grained networking policies within GCP, or between GCP and other public clouds or on-premises data. [Documentation](#).

The Case for Using Informatica Intelligent Cloud Services and Google BigQuery for Cloud Analytics

An agile organization needs to react quickly in response to evolving circumstances, change, and emerging data-driven insights. And these business changes—such as launching new products or services or building new applications to support those products and services—spur the need for new analytics to surface a more robust understanding of those products and services.

Among the new challenges the IT or data integration/analytics organization should expect are new requirements to support and integrate new data sources, manage new data transformations, and extend the data warehouse to accommodate new sets of data. Automating these kinds of tasks with IICS and BigQuery mitigates many of the bottlenecks arising from:

- **Custom code**: Independently building complete solutions for intelligent analytics requires expertise and resources. A complete custom solution needs to connect existing and new data sources, perform transformations, schedule and monitor jobs, and deliver notifications. It will need ongoing revisions whenever SaaS vendors update their application’s APIs. Resources are also needed to build, test, and maintain the solution.

- **Siloed SaaS applications**: Chances are various departments in your organization are deploying multiple SaaS applications such as Salesforce, Marketo, or NetSuite (and maybe even multiple instances of some of these). To get a holistic view of your customers, products, and services, you need to integrate all of these disconnected data sources.

- **On-premises applications**: Accessing and combining data from both SaaS and traditional applications behind the firewall introduces challenges and complexities. On-premises applications typically have their own proprietary APIs or communications protocol. For example, SAP has a web services interface with options for interfacing via tables and proprietary, intermediate document (IDoc) and BAPI.

- **Security considerations**: Protecting sensitive data about your customers and their product and service usage, as well as your internal data, is always critical. Infrastructure security considerations include data transmission, data standards and connectivity, data governance, and audit compliance.
Informatica Intelligent Cloud Services

Informatica Intelligent Cloud Services is a robust data management and integration solution that provides the means for your cloud and on-premises applications to coexist in a powerful and easily managed partnership. With IICS, your data is readily available wherever it resides—in the cloud or on-premises—reliably meeting your organization’s security and compliance standards.

The IICS web-based visual designer makes it easy to create transformations and mappings. Connectors simplify integration to hundreds of SaaS applications, on-premises applications, databases, and enterprise data warehouse options. Informatica continuously updates these integrations, enabling you to focus your time and resources on your business. All of the transformations are supported out-of-the-box, as are functions to transform data sources into enterprise data warehouse targets. IICS lets you schedule data integration jobs, monitor progress, and send notifications when jobs are completed or failed.

Informatica Intelligent Cloud Services complies with security requirements for data centers, networks, systems, databases, and IICS itself. Informatica Intelligent Cloud Services only stores metadata required to perform data integration tasks; no data ever passes through or resides on Informatica servers. More details about security functionality is available at this web page.

Informatica Intelligent Cloud Services components include:
• A six-step synchronization task wizard
• Mapping Designer
• Secure Agent
• Extensive connectivity options

Six-Step Synchronization Task Wizard

Informatica Intelligent Cloud Services lets you quickly create a basic synchronization task using a six-step wizard. You simply select a data source with its objects, specify a target and map the source fields to the target object fields, set filters on incoming records, indicate data transformations using expressions with hundreds of functions, do a lookup for complementary data from another data source, schedule the task, and set up notifications.

![Figure 1: IICS Six-Step Synchronization Task Wizard](image-url)
Mapping Designer
For more complex tasks, IICS provides an intuitive visual design tool called Mapping Designer. It lets you visually design mappings and transformations by dragging and dropping elements of a task. This might include ingesting your data from multiple sources; using normalizer, expression, sorting, joiner, or aggregation transformations; and targeting multiple databases or applications. Mapping Designer streamlines this process without the need to write any code.

Figure 2: IICS Mapping Designer

Secure Agent
Secure Agent is deployed on-premises or on Google Compute Engine. Data integration transpires among the Secure Agent, data sources, and data targets. Data stays within these three components and does not go to IICS, which only stores the metadata of your data integration tasks and mappings.

Figure 3: Informatica Intelligent Cloud Services Secure Agent
Connectivity Options

The complex details of handling an ever-changing array of data sources and targets are maintained by Informatica. Informatica ensures that the connectivity for every data source and target are up-to-date in each release of IICS, and connectivity to new sources and targets are regularly added.

IICS provides connectivity to Google Cloud Platform for the following services:

- Google BigQuery
- Google Cloud SQL (Supports MySQL and PostgreSQL)
- Google Cloud Storage
- Google Cloud Spanner

Google BigQuery

Informatica offers the scale and performance to handle the growing needs of an organization. Informatica Intelligent Cloud Services processes more than 2 trillion records each month for more than 7,000 customers, connecting 100,000 applications, databases, and other endpoints. Coupled with BigQuery’s high-performance data warehouse, the result is a blazingly fast, cloud analytics environment.

BigQuery is a fully managed, petabyte-scale, low cost enterprise data warehouse for analytics. As a serverless data warehouse, there is no infrastructure to manage. This eliminates the need for a database administrator, so you can focus on analyzing data to find meaningful insights using a familiar SQL interface. BigQuery is a powerful big data analytics platform used by all types of organizations, from startups to Fortune 500 companies.

BigQuery separates the functionality of storage and compute systems, enabling you to scale and pay for each according to your needs. Using the processing power of Google’s infrastructure, BigQuery enables extremely fast SQL queries, scanning a terabyte in seconds and a petabyte in minutes.
To make your data available globally with low latency, BigQuery storage automatically replicates data across regions. It also manages the technical aspects of storing your structured data, including compression, encryption, replication, performance tuning, and scaling. BigQuery stores data in the Capacitor columnar data format and offers the standard database concepts of tables, partitions, columns, and rows.

**BigQuery High-Level Decomposition: Data-Center Scale Distributed Query**

![Figure 5: Google BigQuery provides an easy-to-use data warehouse for analytics](image)

BigQuery compute uses a dynamic serving tree, which automatically sizes the computing units required based on each query. By default, 2,000 virtual CPUs are spun up for each query. Virtual CPUs are added automatically when queries require more resources.

When you choose BigQuery as a target, IICS automatically manages a number of complex operations for you. Behind the scenes, IICS optimizes the load, stages data files to Cloud Storage on your behalf, and issues commands to complete the loading process into BigQuery. The following diagram depicts the load process steps:

![Figure 6: Loading Process into Google BigQuery](image)
Google Cloud Platform
Google Cloud Platform lets you use Compute Engine to host IICS Secure Agent and VPC to set up your network, routing, and firewall. Alternatively, Cloud Interconnect can connect your infrastructure to VPC networks directly or through an encrypted tunnel.

Cloud Storage
Cloud Storage is unified object storage for developers and enterprises, supporting live data serving, data analytics, machine learning, and data archiving. Cloud Storage offers four storage classes: Multi-Regional, Regional, Nearline, and Coldline Storage. All storage classes offer the same throughput, low latency (time to first byte is typically tens of milliseconds), and high durability. The classes differ by their availability, minimum storage durations, and pricing for storage and access. Multi-Regional Storage, which maintains your data in geographically distinct locations, is well-suited for objects that are frequently accessed from locations around the world, such as website content. Regional Storage is appropriate when you need to access the data only within a narrow geographic region (e.g., in the same location as your Compute Engine instances that use it for analysis). For data that you do not access frequently, Nearline Storage provides a low-cost option. For backup archiving for disaster recovery, Coldline Storage offers a cost-effective solution.

By default, Google Cloud Storage manages server-side encryption keys on your behalf and encrypts user data at rest using AES-256. Each encryption key is itself encrypted, with a regularly rotated set of master keys. There is no setup or configuration required, no need to modify the way you access the service, and no visible performance impact. Data is automatically and transparently decrypted when read by an authorized user.

Common Use Cases
Here are three common use cases that describe how most organizations can benefit from IICS.

Born in the Cloud: Organizations that use the entire technology stack offered in the cloud (including SaaS, PaaS, and IaaS).

These organizations typically do not own any on-premises warehousing hardware. Instead, they use Google Compute Engine virtual machines. They also do not have traditional databases (i.e., they opt to use Google Cloud SQL, instead) or on-premises CRM systems (i.e., they use a SaaS application like Salesforce).

This type of organization also prefers to avoid investing in the hardware needed to set up an on-premises data warehouse. Instead, they use cloud-based data warehouses, such as Google BigQuery.

Informatica Intelligent Cloud Services plugs into this type of architecture seamlessly using the cloud-hosted Secure Agent or deploying the Secure Agent on Google Compute Engine.
Figure 7: Born in the Cloud Architecture

**Extend into Cloud**: Organizations that have invested in an on-premises technology stack, but want to avoid additional investment, prefer to leverage cloud technology to extend or enhance their on-premises investments.

This approach reduces both cost and implementation time required. By leveraging Google BigQuery, on-premises tasks, like procuring hardware and customizing data warehousing, become unnecessary. Because Google BigQuery is available immediately after signing up for it, lead time is reduced. Using IICS to import data into BigQuery empowers you to start generating analytics within minutes without requiring updates to your on-premises technology stack.

Figure 8: Extend into Cloud Architecture

*“Lift and Shift” into Cloud*: Organizations are moving their technology stack to the cloud—lifting the data and shifting it to new endpoints or data stores—to reduce operating expenses.
Rather than merely extending their data warehousing into the cloud, many organizations are fully migrating data warehousing and other technologies into the cloud. Informatica Intelligent Cloud Services enables easy data warehouse migration to Google BigQuery by allowing you to reuse and redirect existing Informatica workflows that are feeding data to other on-premises or cloud data warehouses, with little modification. Similarly, Informatica PowerCenter workflows can also be easily migrated to IICS or be modified to load data into Google Cloud.

**Deployment Steps**

**Informatica Intelligent Cloud Services Setup**

Log into IICS by accessing this link. If you do not have an account, click the *Don’t have an account?* link on the login page to create a 30-day free trial account.

**Secure Agent**

Informatica Intelligent Cloud Services requires a Secure Agent, which does the actual data processing work. This small-footprint agent is typically installed locally on a server within your organization’s network. Informatica can also provide a hosted Secure Agent option running inside the IICS environment.

A Secure Agent installed on your local server is preferable with any of the following scenarios:
1. You need to access on-premises systems such as database, data warehouse, applications, or files (e.g., CSVs) on your local server.
2. Your organization opts out of processing data in an Informatica-hosted agent environment.
3. The connectors you want to use are not supported by the Informatica-hosted agent.

Otherwise, it is recommended that you use the Informatica-hosted Secure Agent to perform data integration between cloud sources and target systems.
Google Compute Engine and Networking

Installation of Secure Agent on Compute Engine is similar to installation on a local server. In conjunction with Compute Engine, you can configure VPC and Cloud VPN to securely connect your on-premises network to GCP.

Secure Agent Group

Secure Agent group allows pooling more than one Secure Agent in a group. (This feature requires a Secure Agent Cluster license.)

This option offers the following functionality:
1. Balance workload across multiple agents and servers. Data integration tasks are dispatched to available agents in a round-robin fashion.
2. For greater availability and performance, the data workloads are distributed among the available Informatica agents. If one of the agents is not available, the other agents in the group still perform data integration tasks.
3. Optimize agent utilization through Secure Agent groups that can be shared with suborganizations.
4. Segregate Secure Agent groups to process tasks from different departments within your organization.

Learn more about Secure Agent Groups.

Once logged in, select the Administrator service option.
Click **Runtime Environments** on the left panel. Click the **Download Secure Agent** button on the top right to download Secure Agent. Secure Agent is available for Linux and Windows operating systems.

After downloading the Secure Agent, proceed to install it on a local server or Compute Engine. In Windows OS, the wizard guides you through the installation, step by step. At the end of the installation process, it prompts you for your IICS login and password. These credentials link the Secure Agent to your IICS org.

Documentation for Secure Agent installation on Windows OS is available [here](#). Documentation for Secure Agent installation on Linux OS is available [here](#).

Confirm that installation has been successful by logging into IICS and navigating to **Administrator > Runtime Environments**. The status of your Secure Agent should be "Up and Running." (Allow a few minutes for the Secure Agent to complete the startup process, as it does an upgrade check when starting up.)

### Data Source and BigQuery Connection Setup

Informatica has hundreds of connectors to enterprise databases, data warehouses, and applications available. Go to **Administrator > Add-On Connectors** to find the connectors you need and click **Free Trial** button to start using it. The connector then shows up in the connector type list when you create a connection. If you do not see the connector you need, please contact Informatica Support, as some connectors may be in private mode.

Note: You should be familiar with GCP to complete the setup in this section.
If you have not set up your GCP, please do so now. You will need to set up the following on GCP:
- Enable APIs for BigQuery API, Cloud Storage, Cloud Storage JSON API
- Create a Service Account and give privileges to BigQuery and Storage, and save the private key that is in JSON file
- Create a data set if it’s a new BigQuery instance
- Create a Cloud Storage bucket for staging bulk processing data

Details needed from GCP are as follows:
- Service account ID
- Service account key (a long cryptic string in JSON file)
- Project ID
- Cloud Storage path (to the storage bucket)
- BigQuery data set ID (data set name)

Select both Google BigQuery and Google Cloud Storage connectors.

### Google BigQuery Connection Setup

To create a connection, go to **Administrator > Connections**, then click the **New Connection** button.
Select Google BigQuery, fill in the properties required, and click Test Connection button to verify that a connection was established.

Figure 12: BigQuery Connection Details
<table>
<thead>
<tr>
<th>Connection Property Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connection Name</td>
<td>Enter connection name. This name shows up when you select source or target connection in integration tasks and mapping.</td>
</tr>
<tr>
<td>Description</td>
<td>Enter tags so you can search for it.</td>
</tr>
<tr>
<td>Type</td>
<td>Select Google BigQuery.</td>
</tr>
<tr>
<td>Runtime environment</td>
<td>Select the Secure Agent that you want to run the tasks.</td>
</tr>
<tr>
<td>Service account ID</td>
<td>Google Cloud Platform service account ID.</td>
</tr>
<tr>
<td>Service account key</td>
<td>Account key is in the JSON file downloaded when you create the service account.</td>
</tr>
<tr>
<td>Project ID</td>
<td>Google Cloud Platform project ID that you want the connector to associate with.</td>
</tr>
<tr>
<td>Storage path</td>
<td>Google Cloud Storage path for staging purpose. Add gs:// prefix in the path.</td>
</tr>
<tr>
<td>Connection mode</td>
<td>Simple, hybrid, or complex</td>
</tr>
<tr>
<td>Schema definition file path</td>
<td>Directory to schema definition file in the selected runtime environment.</td>
</tr>
<tr>
<td>Data set ID</td>
<td>Google BigQuery data set name.</td>
</tr>
</tbody>
</table>

Table 1: Connection Names and Descriptions
Google Cloud Storage Connection Setup
Select Google Cloud Storage, fill in the properties required, and click the Test Connection button to verify that a connection can be made.

![Google Cloud Storage Connection Setup](image)

Figure 13: Google Cloud Storage Connection Details

<table>
<thead>
<tr>
<th>Connection Property Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connection Name</td>
<td>Enter connection name. This name shows up when you select source or target connection in integration tasks and mapping.</td>
</tr>
<tr>
<td>Description</td>
<td>Enter tags so you can search for it.</td>
</tr>
<tr>
<td>Type</td>
<td>Select Google Cloud Storage.</td>
</tr>
<tr>
<td>Runtime environment</td>
<td>Select the Secure Agent that you want to run the tasks.</td>
</tr>
<tr>
<td>Service account ID</td>
<td>Google Cloud Platform service account ID.</td>
</tr>
<tr>
<td>Service account key</td>
<td>Account key is in the JSON file downloaded when you create the service account.</td>
</tr>
<tr>
<td>Project ID</td>
<td>Google Cloud Platform project ID that you want the connector to associate with.</td>
</tr>
<tr>
<td>File path in Google Cloud</td>
<td>Google Cloud Storage path for staging purpose. gs:// prefix is NOT needed.</td>
</tr>
</tbody>
</table>

Table 2: Connection Property Names and Descriptions
Data Integration Task

The next step is to configure a data integration task using the six-step wizard. Go to Data Integration service, click New, select Synchronization Task and click the Create button.

Figure 14: Configuring a Data Integration Task

1. Enter the Task Name and select Insert for Task Operation to insert into Target. Other operations available are Update, Upsert, and Delete on the Target side. Click Next.

Figure 15: Synchronization Task Definition
2. Select the data source from the Connection dropdown, or define a new connection from this screen (click New button if that is the case). Select the Source Type. Select either a Single or Multiple table(s) or object(s). When Multiple is selected, either a relationship that is known by Informatica will be prompted or a relationship between tables/objects can be established manually. Performance is better using this method as the join is performed in the source database or application. A custom Saved Query can also be used to select the source data.

![New Insert into BigQuery](image)

**Source Details**
- Connection: SQL Server
- Source Type: Single
- Source Object: TPCH_Customer

**Data Preview**

<table>
<thead>
<tr>
<th>c_custkey</th>
<th>c_name</th>
<th>c_address</th>
<th>c_nationkey</th>
<th>c_phone</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Customer#000000001</td>
<td>1</td>
<td>15</td>
<td>25-989-741-2988</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>Customer#000000002</td>
<td>2</td>
<td>13</td>
<td>23-768-687-3665</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>Customer#000000003</td>
<td>3</td>
<td>1</td>
<td>11-719-748-3364</td>
<td>...</td>
</tr>
<tr>
<td>4</td>
<td>Customer#000000004</td>
<td>4</td>
<td>14</td>
<td>128-190-5944</td>
<td>...</td>
</tr>
<tr>
<td>5</td>
<td>Customer#000000005</td>
<td>5</td>
<td>3</td>
<td>13-750-942-6364</td>
<td>...</td>
</tr>
<tr>
<td>6</td>
<td>Customer#000000006</td>
<td>6</td>
<td>20</td>
<td>30-114-968-4931</td>
<td>...</td>
</tr>
<tr>
<td>7</td>
<td>Customer#000000007</td>
<td>7</td>
<td>18</td>
<td>28-190-982-9739</td>
<td>...</td>
</tr>
<tr>
<td>8</td>
<td>Customer#000000008</td>
<td>8</td>
<td>17</td>
<td>27-147-574-9335</td>
<td>...</td>
</tr>
<tr>
<td>9</td>
<td>Customer#000000009</td>
<td>9</td>
<td>8</td>
<td>18-338-906-3675</td>
<td>...</td>
</tr>
<tr>
<td>10</td>
<td>Customer#000000010</td>
<td>10</td>
<td>5</td>
<td>15-741-346-9870</td>
<td>...</td>
</tr>
</tbody>
</table>

Figure 16: SQL Server Data Source and Data Preview
3. Select the **Google BigQuery** target from the **Connection** dropdown. If you don’t see an option presented for the table that you want to insert into, click **Create Target** button to create it, as seen in the screenshot.

Note: If you experience an error because you are referring to no target object and your BigQuery data set does not have any tables, create a dummy table in the BigQuery data set to resolve this problem.

**New Insert into BigQuery**

<table>
<thead>
<tr>
<th>Target Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connection: <em>Google BigQuery - simple</em></td>
</tr>
<tr>
<td>Target Object: <em>TPCH_Customer</em></td>
</tr>
<tr>
<td>Child Object: <em>None Found</em></td>
</tr>
</tbody>
</table>

**Data Preview**

<table>
<thead>
<tr>
<th>c_custkey</th>
<th>c_name</th>
<th>c_address</th>
<th>c_nationkey</th>
<th>c_phone</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>7948</td>
<td>Customer#0000007948</td>
<td>KYVo3igYlC,dbZ6m,9xw21f3zE5snTIN...</td>
<td>0</td>
<td>10-981-358-8909</td>
<td>...</td>
</tr>
<tr>
<td>12512</td>
<td>Customer#000012512</td>
<td>5c3wZFAbgk5uV7C66Nn</td>
<td>0</td>
<td>10-406-287-7724</td>
<td>...</td>
</tr>
<tr>
<td>3169</td>
<td>Customer#000003169</td>
<td>gld8F4FV9Ac9Waf</td>
<td>0</td>
<td>10-471-622-4713</td>
<td>...</td>
</tr>
<tr>
<td>48</td>
<td>Customer#000000048</td>
<td>0UU,PkBpFwemNB</td>
<td>0</td>
<td>10-508-348-5882</td>
<td>...</td>
</tr>
<tr>
<td>5503</td>
<td>Customer#000005503</td>
<td>D25aWvkhVY</td>
<td>0</td>
<td>10-902-385-3078</td>
<td>...</td>
</tr>
<tr>
<td>13102</td>
<td>Customer#0000013102</td>
<td>Sw0g,63p,8dqKWh6QIjFCF</td>
<td>0</td>
<td>10-668-392-6796</td>
<td>...</td>
</tr>
<tr>
<td>13378</td>
<td>Customer#000013378</td>
<td>651UdpvV3Spv3gbgNn,ZHUFhmdnc...</td>
<td>0</td>
<td>10-289-753-6505</td>
<td>...</td>
</tr>
<tr>
<td>8096</td>
<td>Customer#000008096</td>
<td>FEAmmc,swjGZ3s7/3jKqGtfV...</td>
<td>0</td>
<td>10-912-605-2020</td>
<td>...</td>
</tr>
<tr>
<td>11349</td>
<td>Customer#000011349</td>
<td>nUS7YDRCHJUPHYtkmQ60NZNzClO</td>
<td>0</td>
<td>10-984-212-9943</td>
<td>...</td>
</tr>
<tr>
<td>793</td>
<td>Customer#00000793</td>
<td>SHK1mp2Kxwrb,A0x3Co1uhcwr</td>
<td>0</td>
<td>10-404-953-9048</td>
<td>...</td>
</tr>
</tbody>
</table>

Figure 17: BigQuery Target and Data Preview of Existing Data
4. Data filters lets you filter out any incoming source data you want to ignore. Click **New** button if you want to add a filter.

**New Insert into BigQuery**

<table>
<thead>
<tr>
<th>Definition</th>
<th>Source</th>
<th>Target</th>
<th>Data Filters</th>
<th>Field Mapping</th>
<th>Schedule</th>
</tr>
</thead>
</table>

Row Limit

- Process all rows
- Process only the first 100 rows

Data Filters: New...

There are no filters defined. The task will process all data from source.

Figure 18: Data Configuration

5. Field Mapping shows the mapping of incoming source fields (on the left side) and how the fields are mapped to the target fields (on the right side). If the source and target fields are the same, the wizard will auto match them. The actions column provides several options. Click the **fx** icon to transform the field data using standard expressions. The magnifying glass lets you perform a lookup to search for a specific field from another data source and retrieve fields to be used in the mapping. To first clear the mapping for a field (before doing a lookup), click the **x** icon.

**New Insert into BigQuery**

<table>
<thead>
<tr>
<th>Source: TPCH_Customer</th>
<th>Target: TPCH_Customer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Edit Types</strong></td>
<td><strong>Edit Types</strong></td>
</tr>
<tr>
<td>Status</td>
<td>Name</td>
</tr>
<tr>
<td>✔</td>
<td>c_outkey</td>
</tr>
<tr>
<td>✔</td>
<td>c_name</td>
</tr>
<tr>
<td>✔</td>
<td>c_address</td>
</tr>
<tr>
<td>✔</td>
<td>c_nationkey</td>
</tr>
<tr>
<td>✔</td>
<td>c_phone</td>
</tr>
<tr>
<td>✔</td>
<td>c_octbol</td>
</tr>
<tr>
<td>✔</td>
<td>c_mktsegment</td>
</tr>
<tr>
<td>✔</td>
<td>c_comment</td>
</tr>
</tbody>
</table>

Figure 19: Source to Target Field Mapping

Example: Transform all characters of c_name field into a capital letter. There are hundreds of functions and operators for transforming the incoming fields.
Figure 20: Expression Transformation to Capitalize c_name Field

Example: Lookup screen

Figure 21: Lookup Transformation - Not Configured/Used
6. This step provides tools to configure a schedule to run the task, trigger an email notification, and specify pre- and post-processing commands and advanced properties for Google BigQuery.

Figure 22: Task Scheduling, Email Notification, and Advanced Options
A schedule can be created, with the option to create a recurring schedule using the repeats option.

**New Schedule**

A schedule specifies when and how often a job runs. Enter a name, start time, and repeat frequency for the schedule.

**Schedule Details**

Schedule Name:*  
Daily at 3am

Description:

**Schedule Options**

Starts:*  01/03/2018  at 03 : 00

Time Zone:  Pacific Daylight Time, Los Angeles

Repeats:  Daily

**Repeat Frequency Options: Daily**

Run the task:

- **Every day**
- **Every week**

Figure 23: Create New Schedule
Scroll down to view *Advanced Target Properties* for a BigQuery target that can be set to further refine the configuration.

### New Insert into BigQuery

<table>
<thead>
<tr>
<th>1 Definition</th>
<th>2 Source</th>
<th>3 Target</th>
<th>4 Data Filters</th>
<th>5 Field Mapping</th>
<th>6 Schedule</th>
</tr>
</thead>
</table>

#### Advanced Target Properties

- **Target Dataset ID:**
- **Target Table Name:**
- **Create Disposition:** Create never
- **Write Disposition:** Write append
- **Write Mode:** Bulk
- **Streaming Template Table Suffix:**
- **Rows per Streaming Request:** 500
- **Staging File Name:**
- **Data format of the staging file:** JSON (Newline Delimited)
- **Persist Staging File After Loading:**
- **Enable Staging File Compression:**
- **Job Poll Interval In Seconds:** 10
- **Number of Threads for Uploading Staging File:** 1
- **Local Stage File Directory:**
- **Allow Quoted Newlines:**
- **Field Delimiter:**

*Figure 24: Advanced Properties for Google BigQuery Target*
<table>
<thead>
<tr>
<th>Advanced Property Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target dataset ID</td>
<td>Optional field used to overwrite the Dataset ID specified in BigQuery connection.</td>
</tr>
<tr>
<td>Target table name</td>
<td>Optional field used to overwrite the target Table Name specified in Synchronization Task.</td>
</tr>
<tr>
<td>Create disposition</td>
<td>If target table does not exist, Create never returns an error. Create if needed creates the table.</td>
</tr>
<tr>
<td>Write disposition</td>
<td>Defines how to insert if the target table already exists. Default Write append, appends into existing table. Write empty inserts if target table is empty, and returns an error otherwise. Write truncate deletes existing records prior to inserting into the target table. Applicable to bulk write mode only.</td>
</tr>
<tr>
<td>Write mode</td>
<td>Default Bulk mode inserts records into target in bulk while Streaming writes row by row, which lets you query data from the target without a delay.</td>
</tr>
<tr>
<td>Streaming template table suffix</td>
<td>Suffix name for the templateSuffix BigQuery parameter. Template table is used when streaming data into BigQuery. It allows you to split a logical table into many smaller tables without adding complex code. (Applicable to streaming write mode only.)</td>
</tr>
<tr>
<td>Rows per streaming request</td>
<td>Number of rows per streaming request. Default is 500 rows; maximum row size is 10MB. (Applicable to streaming write mode only.)</td>
</tr>
<tr>
<td>Staging file name</td>
<td>Use to overwrite the default file name. Default file name is Informatica_Stage_File_[unix timestamp]. [data format: JSON, CSV]. (Applicable to bulk write mode only.)</td>
</tr>
<tr>
<td>Data format of the staging file</td>
<td>Data format of Cloud Storage staging files. JSON – supports nested and repeated fields or CSV formats.</td>
</tr>
<tr>
<td>Persist staging file after loading</td>
<td>This option allows the staging file to be persisted in Google Cloud Storage. Staging files are deleted by default. (Applicable to bulk write mode only.)</td>
</tr>
<tr>
<td>Enable staging file compression</td>
<td>Compress the staging file to reduce cost of storage and transfer time from Secure Agent to Google Cloud Storage.</td>
</tr>
<tr>
<td>Job poll interval in seconds</td>
<td>Number of seconds to poll for status of job. Default is 10 seconds.</td>
</tr>
<tr>
<td>Number of threads for uploading staging file</td>
<td>If increased from default of 1, increase Secure Agent’s DTM JVMOption3 heap size as well. (Applicable to bulk write mode only.)</td>
</tr>
</tbody>
</table>
Local stage file directory

If using local Secure Agent, this directory is used to store temporary files prior to writing to Cloud Storage staging file. (Applicable to bulk write mode only.)

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allow quoted newlines</td>
<td>Allows newline character inside quotes in a single record in a CSV file.</td>
</tr>
<tr>
<td>Field delimiter</td>
<td>Default field delimiter is comma (,).</td>
</tr>
<tr>
<td>Allow jagged rows</td>
<td>Allows rows without trailing columns in a CSV file. Default is to not allow this.</td>
</tr>
<tr>
<td>Success file directory</td>
<td>Not applicable.</td>
</tr>
<tr>
<td>Error file directory</td>
<td>Not applicable.</td>
</tr>
</tbody>
</table>

Table 3: Advanced Property Names and Descriptions

**Mapping Designer**

Mapping Designer offers an easy method for accommodating situations with more complex requirements. Mapping Designer includes many more built-in transformations than just expression and lookup. It also supports mappings with multiple data sources and targets.

![Mapping Designer](image)

Figure 25: Mapping Designer
Supported transformations are: source, target, joiner, sequence, filter, sorter, router, expression, lookup, web services, normalizer, hierarchy builder, hierarchy parser, mapplet, aggregator, union, and SQL.

To create a new mapping, go to **New**, click **Mappings**. Select **Mappings** and click the **Create** button.

![New Asset](image)

**Figure 26: Create New Mapping**

Select transformation from the available transformations, drag and drop onto the canvas, and fill in transformation properties at the bottom panel.

Click the checkmark icon at the top right to validate the mapping. This will bring the validation window, which shows transformations with errors in them. If no errors are found, the validation panel displays “Mapping is valid.”

![Mapping Validation Panel](image)

**Figure 27: Mapping Validation Panel on the Right**
Once the mapping is validated, click the **Save** button to save it. Then click the vertical ellipsis icon in the upper right, and select New Mapping Task.

**Figure 28: Create New Mapping Task**

Fill in **Task Name** and select **Runtime Environment**, then click **Next**.

**Figure 29: Select Runtime Environment for Mapping Task**
Choose a schedule if needed, then click **Finish**.

Figure 30: Configure Schedule

Click **Run** to run the mapping.

Figure 31: Mapping Task

The mapping configuration is queued for execution.
Advanced Topics

BigQuery RECORD Data Type Support

BigQuery RECORD data type is supported in IICS. Use HierarchyParser transformation to read RECORD data type from BigQuery, and use HierarchyBuilder transformation to write into a BigQuery table that has a RECORD data type.

Figure 32: Use HierarchyBuilder Transformation to Insert into RECORD Data Type

Do the following to insert into a BigQuery target table that has RECORD data type:

1. Sort the secondary source table based on the foreign key. In source table property, go to Source tab, then expand Query Options. Click Configure for the sort option and select the sort key.
2. Link primary and secondary table into HierarchyBuilder transformation.

3. In HierarchyBuilder transformation, go to Output Settings tab and load the JSON schema representing the BigQuery target table.
Figure 34: JSON Schema for BigQuery Target Table
4. Go to the Field Mapping tab and mark primary key and foreign key fields. Then map the fields on the left to the fields on the right.

Figure 35: HierarchyBuilder Transformation: Mapping Relational to Hierarchy Field, Primary and Foreign Keys Configured
5. In BigQuery target, select target table, then go to Field Mapping. Map the field Output on the left to the STRING_DATA on the right.

![Field Mapping to BigQuery Target](image1)

6. Here's the target BigQuery table preview.

![BigQuery Table with RECORD Data Type](image2)
Informatica PowerCenter Customers
Informatica PowerCenter is the on-premises alternative to IICS data integration capabilities. Informatica's goal is for IICS to offer feature parity with PowerCenter in the near future. However, thousands of customers using PowerCenter have built extensive data pipelines/mappings to support their enterprise data warehouse and other data integration requirements. For customers relying on existing PowerCenter resources, but still trying to extend and/or migrate their on-premises data warehouse to a cloud data warehouse, there are two options: migrate from PowerCenter to IICS or use PowerExchange® for Google BigQuery to integrate to BigQuery. Both approaches take advantage of Google BigQuery cloud data warehouse.

Migrate from PowerCenter to IICS
This is the easiest option if you do not have a lot of data pipeline/mappings and the transformations in those mappings are supported by IICS. Follow the IICS deployment steps describe above. Informatica Professional Services also offers assistance to convert a large number of PowerCenter mappings to IICS.

PowerExchange for Google BigQuery
This option is best for those that want to continue to use PowerCenter through that transition. This is also appropriate if IICS does not yet support the transformations used in your data pipeline/mappings. PowerExchange for Google BigQuery is a native connector with read/write capabilities and support for advance configurations.
1. In PowerCenter *Target Designer*, select *Create PowerExchange for BigQuery Target*.
2. Select connection to Google BigQuery if there are existing connections.

![Available Connections](image1)

**Figure 39: Available Connections**

3. Click *Import from a new connection* to create a new connection. See above Google BigQuery Connection Setup section.

![Create a Google BigQuery Connection](image2)

**Figure 40: Create a Google BigQuery Connection**

4. Click **Test** to test the connection.

![Create Connection Successful Message](image1.png)

Figure 41: Create Connection Successful Message

5. Click **OK** and click **Save Connection**. Enter the connection name and click **Done**.

![Name the Connection](image2.png)

Figure 42: Name the Connection
6. With the new connection selected, Click Next.

Figure 43: Select Connection

7. Available data sets are displayed. Select a data set and select a table. Click Import.

Figure 44: Select Data Set and Table
Summary

Unleashing the disruptive power of data brings huge opportunities for all organizations—along with headaches if you don’t choose and use the right tools and architecture. With new ways to intelligently leverage data, new business opportunities—such as optimally monetizing products and services or iterating your offerings—can be pursued and analyzed more quickly and efficiently. However, if your IT environment is not set up to provide the agile support you’ll need to unleash this data, the challenge may prove daunting. Fortunately, Informatica and Google have platforms to support your data-driven digital transformation journey.

The Informatica Intelligent Cloud Services data management and integration solution enables you to integrate hundreds of applications and databases (both on-premises and in the cloud) with Google BigQuery. It is easy to use, allowing marketing and other LOB personnel to build cloud analytics without requiring much IT involvement. The IICS Six-Step Wizard, for example, simplifies and accelerates the process of synchronizing data to start your analytics. Mapping Designer streamlines the creation of complex transformations involving multiple data sources and targets.
Because the GCP technology stack offers low-cost and high-performance services, it is well suited for a broad range of use cases. It offers options to host your IICS Secure Agents on Google Compute Engine, use Google Networking to connect to your on-premises network, store products and service offerings data in Google BigQuery, and archive your data in Google Cloud Storage. Informatica PowerCenter users can leverage PowerExchange for Google BigQuery to connect to BigQuery and maintain the investments they have made in PowerCenter data pipelines/mappings.

Informatica Intelligent Cloud Services also connects to hundreds of on-premises and cloud applications, and data stores. With IICS, you can move to the cloud at your pace, as your organization’s data center of gravity moves to the cloud. IICS releases new data management capabilities and new connectors in every release. For the latest, check out https://www.informatica.com/cloud.