Transferring data to and from Azure

02/12/2018 • 7 minutes to read • Contributors all

In this article

There are several options for transferring data to and from Azure, depending on your needs.

Physical transfer

Using physical hardware to transfer data to Azure is a good option when:

- Your network is slow or unreliable.
- Getting additional network bandwidth is cost-prohibitive.
- Security or organizational policies do not allow outbound connections when dealing with sensitive data.

If your primary concern is how long it will take to transfer your data, you may want to run a test to verify whether network transfer is actually slower than physical transport.

There are two main options for physically transporting data to Azure:

- **Azure Import/Export**. The Azure Import/Export service lets you securely transfer large amounts of data to Azure Blob Storage or Azure Files by shipping internal SATA HDDs or SDDs to an Azure datacenter. You can also use this service to transfer data from Azure Storage to hard disk drives and have these shipped to you for loading on-premises.

- **Azure Data Box**. Azure Data Box is a Microsoft-provided appliance that works much like the Azure Import/Export service. Microsoft ships you a proprietary, secure, and tamper-resistant transfer appliance and handles the end-to-end logistics, which you can track through the portal. One benefit of the Azure Data Box service is ease of use. You don't need to purchase several hard drives, prepare them, and transfer files to each one. Azure Data Box is supported by a number of industry-leading Azure partners to make it easier to seamlessly leverage offline transport to the cloud from their products.

Command line tools and APIs

Consider these options when you want scripted and programmatic data transfer.

- **Azure CLI**. The Azure CLI is a cross-platform tool that allows you to manage Azure services and upload data to Azure Storage.

- **AzCopy**. Use AzCopy from a Windows or Linux command-line to easily copy data to and from Azure Blob, File, and Table storage with optimal performance. AzCopy supports concurrency and parallelism, and the ability to resume copy operations when interrupted. You can also leverage AzCopy to copy data from AWS to Azure. For programmatic access, the Microsoft Azure Storage Data Movement Library is the core framework that powers AzCopy. It is provided as a .NET Core library.

- **PowerShell**. The Start-AzureStorageBlobCopy PowerShell cmdlet is an option for Windows administrators who are used to PowerShell.

- **AdlCopy**. AdlCopy enables you to copy data from Azure Storage Blobs into Data Lake Store. It can also be used to copy data between two Azure Data Lake Store accounts. However, it cannot be used to copy data from Data Lake Store to Storage Blobs.

- **Distcp**. If you have an HDInsight cluster with access to Data Lake Store, you can use Hadoop ecosystem tools like Distcp to copy data to and from an HDInsight cluster storage (WASB) into a Data Lake Store account.

- **Sqoop**. Sqoop is an Apache project and part of the Hadoop ecosystem. It comes preinstalled on all HDInsight clusters. It allows data transfer between an HDInsight cluster and relational databases such as SQL, Oracle, MySQL, and so on. Sqoop is a collection of related tools, including import and export. Sqoop works with HDInsight clusters using either Azure Storage blobs or Data Lake Store attached storage.

- **PolyBase**. PolyBase is a technology that accesses data outside of the database through the T-SQL language. In SQL Server 2016, it allows you to run queries on external data in Hadoop or to import/export data from Azure Blob Storage. In Azure SQL Data Warehouse, you can import/export data from Azure Blob Storage and Azure Data Lake Store. Currently, PolyBase is the fastest method of importing data into SQL Data Warehouse.

- **Hadoop command line**. When you have data that resides on an HDInsight cluster head node, you can use the `hadoop -copyFromLocal` command to copy that data to your cluster's attached storage, such as Azure Storage blob or Azure Data Lake Store. In order to use the Hadoop command, you must first connect to the head node. Once connected, you can upload a file to storage.

Graphical interface

Consider the following options if you are only transferring a few files or data objects and don't need to automate the process.

- **Azure Storage Explorer**. Azure Storage Explorer is a cross-platform tool that lets you manage the contents of your Azure storage accounts. It allows you to upload, download, and manage blobs, files, queues, tables, and Azure Cosmos DB entities. Use it with Blob storage to manage blobs and folders, as well as upload and download blobs between your local file system and Blob storage, or between storage accounts.

- **Azure portal**. Both Blob storage and Data Lake Store provide a web-based interface for exploring files and uploading new files one at a time. This is a good option if you do not want to install any tools or issue commands to quickly explore your files, or to simply upload a handful of new ones.

Data pipeline

Azure Data Factory. Azure Data Factory is a managed service best suited for regularly transferring files between a number of Azure services, on-premises, or a combination of the two. Using Azure Data Factory, you can create and schedule data-driven workflows (called pipelines) that ingest data from disparate data stores. It can process and transform the data by using compute services such as Azure HDInsight Hadoop, Spark, Azure Data Lake Analytics, and Azure Machine Learning. Create data-driven workflows for orchestrating and automating data movement and data transformation.

Key Selection Criteria

For data transfer scenarios, choose the appropriate system for your needs by answering these questions:

- Do you need to transfer very large amounts of data, where doing so over an Internet connection would take too long, be unreliable, or too expensive? If yes, consider physical transfer.

- Do you prefer to script your data transfer tasks, so they are reusable? If so, select one of the command line options or Azure Data Factory.
- Do you need to transfer a very large amount of data over a network connection? If so, select an option that is optimized for big data.
- Do you need to transfer data to or from a relational database? If yes, choose an option that supports one or more relational databases. Note that some of these options also require a Hadoop cluster.
- Do you need an automated data pipeline or workflow orchestration? If yes, consider Azure Data Factory.

Capability matrix

The following tables summarize the key differences in capabilities.

Physical transfer

Capability	Azure Import/Export service	Azure Data Box
Form factor	Internal SATA HDDs or SDDs	Secure, tamper-proof, single hardware appliance
Microsoft manages shipping logistics	No	Yes
Integrates with partner products	No	Yes
Custom appliance	No	Yes

Command line tools

Hadoop/HDInsight

Capability	Distcp	Sqoop	Hadoop CLI
Optimized for big data	Yes	Yes	Yes
Copy to relational database	No	Yes	No
Copy from relational database	No	Yes	No
Copy to Blob storage	Yes	Yes	Yes
Copy from Blob storage	Yes	Yes	No
Copy to Data Lake Store	Yes	Yes	Yes
Copy from Data Lake Store	Yes	Yes	No

Other

Capability	Azure CLI	AzCopy	PowerShell	AdlCopy	PolyBase
Compatible platforms	Linux, OS X, Windows	Linux, Windows	Windows	Linux, OS X, Windows	SQL Server, Azure SQL Data Warehouse
Optimized for big data	No	No	No	Yes [1]	Yes [2]

Capability	Azure CLI	AzCopy	PowerShell	AdlCopy	PolyBase
Copy to relational database	No	No	No	No	Yes
Copy from relational database	No	No	No	No	Yes
Copy to Blob storage	Yes	Yes	Yes	No	Yes
Copy from Blob storage	Yes	Yes	Yes	Yes	Yes
Copy to Data Lake Store	No	No	Yes	Yes	Yes
Copy from Data Lake Store	No	No	Yes	Yes	Yes

[1] AdlCopy is optimized for transferring big data when used with a Data Lake Analytics account.

[2] PolyBase performance can be increased by pushing computation to Hadoop and using PolyBase scale-out groups to enable parallel data transfer between SQL Server instances and Hadoop nodes.

Graphical interface and Azure Data Factory

Capability	Azure Storage Explorer	Azure portal *	Azure Data Factory
Optimized for big data	No	No	Yes
Copy to relational database	No	No	Yes
Copy from relational database	No	No	Yes
Copy to Blob storage	Yes	No	Yes
Copy from Blob storage	Yes	No	Yes
Copy to Data Lake Store	No	No	Yes
Copy from Data Lake Store	No	No	Yes
Upload to Blob storage	Yes	Yes	Yes
Upload to Data Lake Store	Yes	Yes	Yes
Orchestrate data transfers	No	No	Yes
Custom data transformations	No	No	Yes
Pricing model	Free	Free	Pay per usage

* Azure portal in this case means using the web-based exploration tools for Blob storage and Data Lake Store.