# Introduction

## Why Azure Data Factory

* Leverage HDInsight resources
* Document DB
* Data born in Azure (Stream Analytics, ML, Facebook data, etc.)
* Storing and sharing large amount of data; less expensive than SQL in Azure or APS/VM/On prem
* Orchestration and monitoring (ask Wee for SSIS integration)

### Comparison with SSIS

* Needs Windows (ISSERVER service) and SQL Server (SSIS Catalog)
* Batch processing and extensive orchestration (master packages)
* Might be limited for large data sets (big data)

## Getting ready

### Azure account setup

Since Azure Data Factory relies on Azure cloud, you’ll need am Azure account to create, manage and interact with your factory (ies).

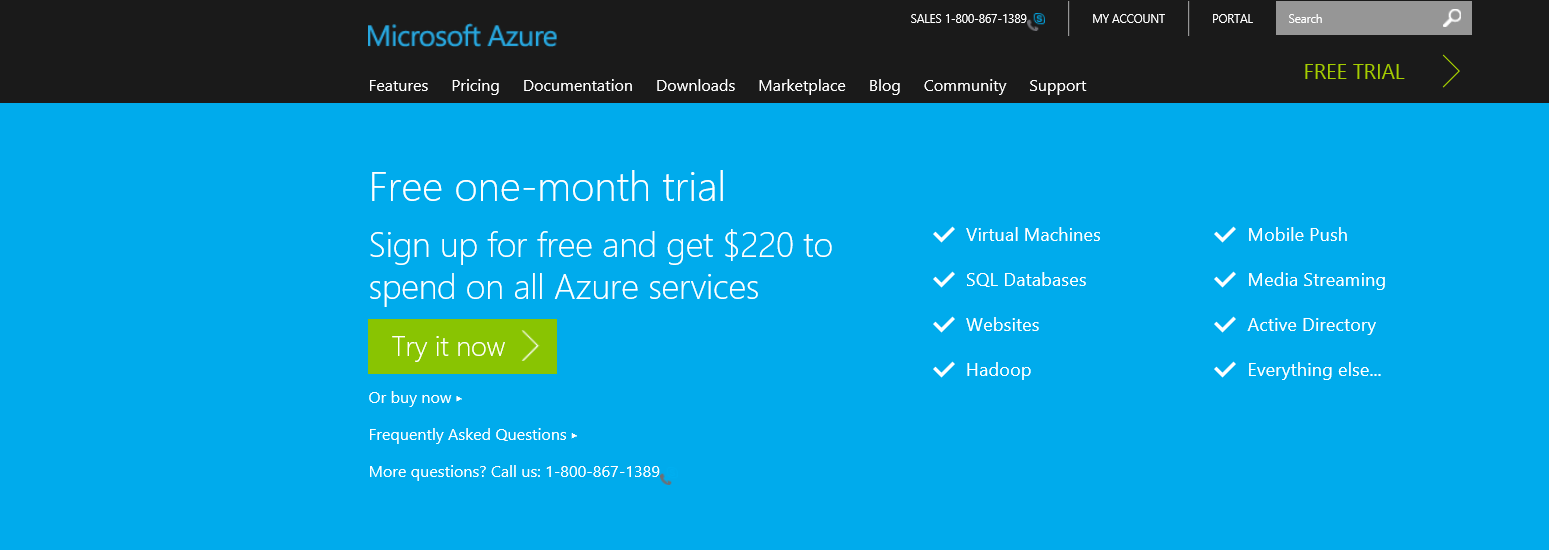
To do so you first go to the [Microsoft Azure web site](http://azure.microsoft.com/en-us/pricing/free-trial/?WT.mc_id=azurebg_US_sem_bing_BR_BRTop_Nontest_FreeTrial_azure&WT.srch=1) and sign in with your Microsoft account that is linked to Microsoft Azure. If you don’t have a Microsoft account, you have the option to get a one-month trial for free.

Figure 1‑1

If you select “Try it now”, you are redirected to the following scree. You now have two options: Sign-in

with your Microsoft account or create a new one by clicking “Sign up now” in the screen below.

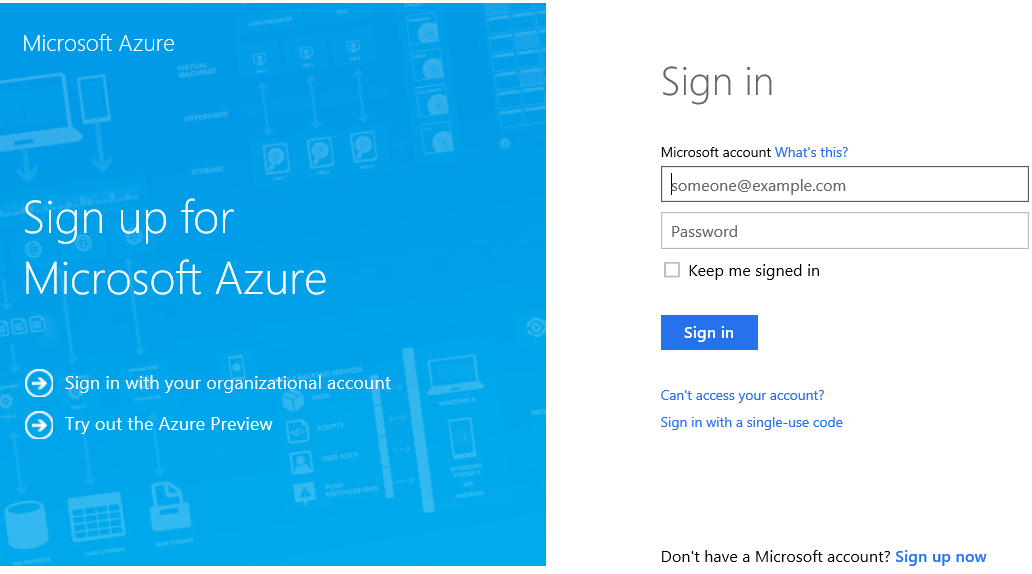


Figure 1‑2

Once you have setup your account, go the new [Azure portal](https://portal.azure.com/) and enter your credentials to log in. You should see a screen similar to the following one:

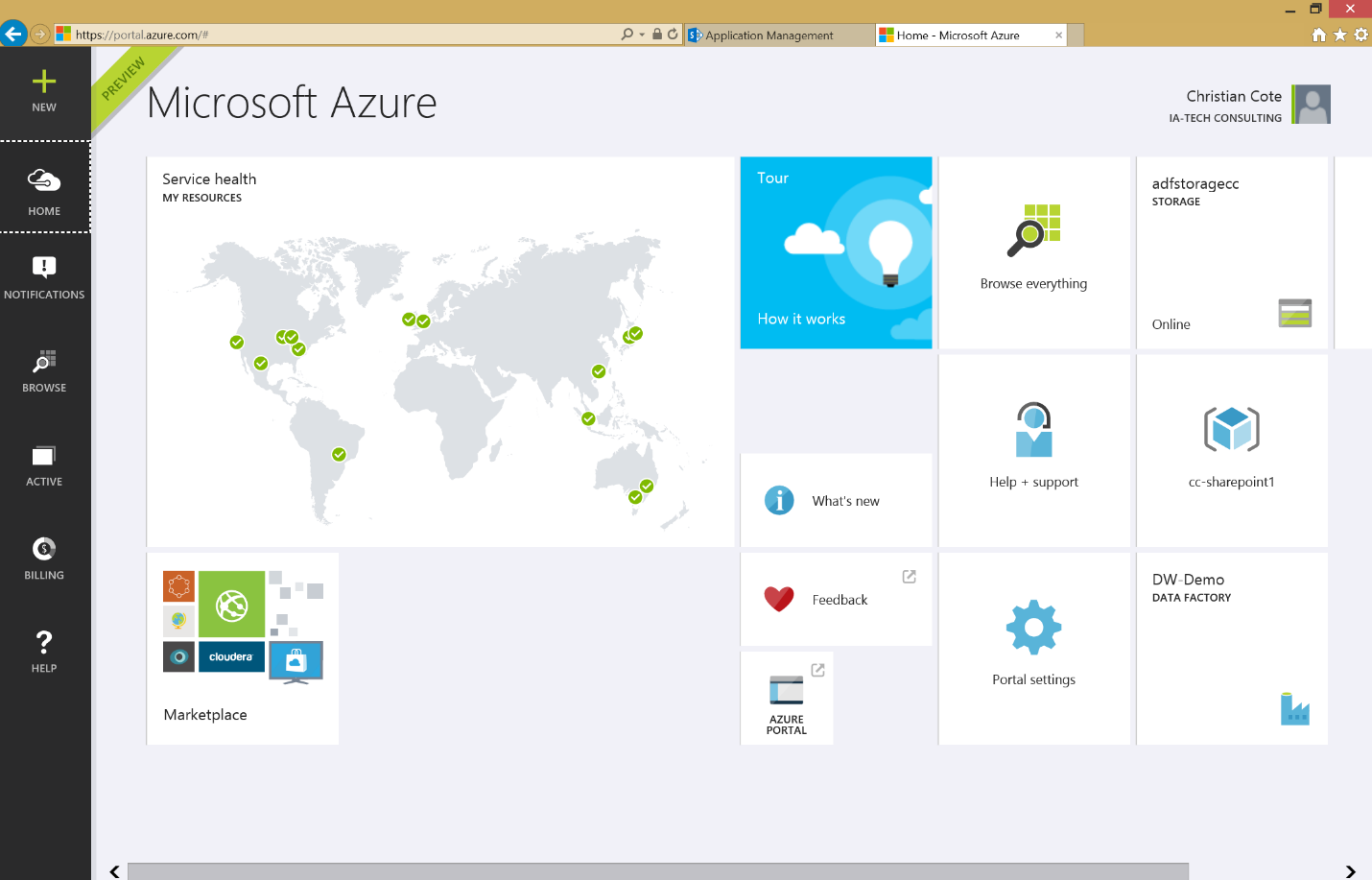




Figure 1‑3

In this book, we’ve changed the default theme to use the dark one. To do this, click on the portal settings (gear). The following screen appears:

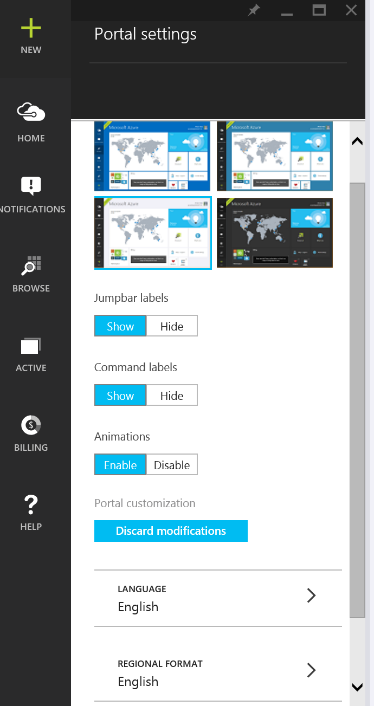


Figure 1‑4

This screen allows you to customize the portal as you like. But the only thing we want here is to use the dark theme. So we click on the dark theme to select it (Figure 1‑5)



Figure 1‑5

The background instantaneously turn black. Now, click on the X at the top of the Portal setting blade to close it as shown in Figure 1‑6 below.



Figure 1‑6

### Creating a storage account

The next step is to configure the storage account. Almost every services in Azure are using storage. The counterpart of Azure storage on premise would be the hard drives or thumb drives on computers or servers. They contains files that are used by us or various services in Azure. Almost everything we do on Azure generate files. For example, when we execute a map reduce job on an HDInsight cluster, it generates a lot of log files that is the result of the job execution as well of the execution by itself.

#### Create a storage from the portal

From the portal main page, click on the + sign, select “Data + Storage” and “Storage” at the right.

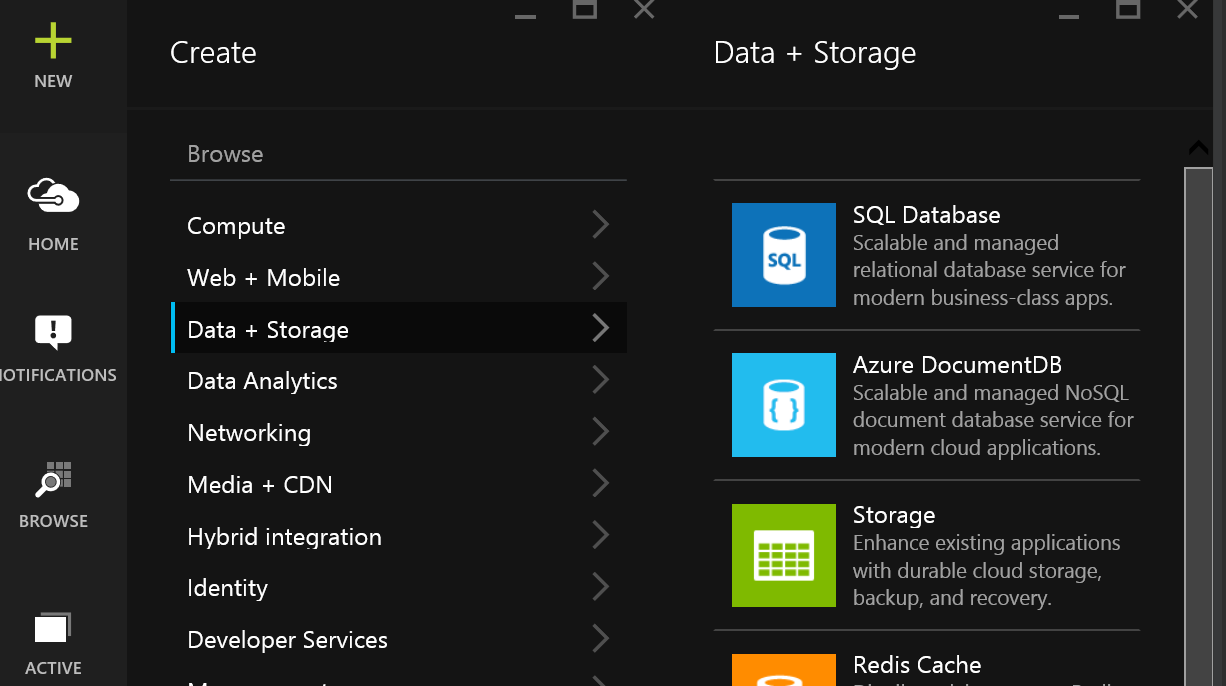


Figure 1‑7

The storage account blade appears.

##### Storage name

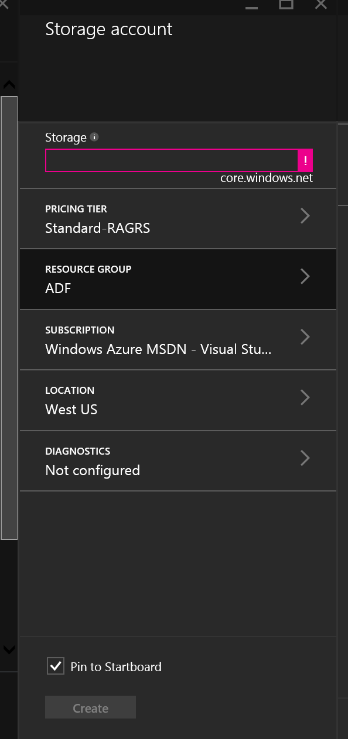


Figure 1‑8

The name of the storage has to be in lowercase, can only contains letters and digits and has to be unique. The reason for the uniqueness is that the storage has an address accessible via internet. For example, if I select MyAdfStorage, it will be known as being myadfstorage.core.windows.net and that’s how many tools or service will refer to it.

As you can see on Figure 1‑8 the name is validated as we enter it. In that case, the text box is magenta because we didn’t provided any name for the storage. Also if you hover you mouse over the exclamation mark at the right of the storage name text box (Figure 1‑9), you’ll get the validation error message that tells you what’s wrong with the name of your storage.

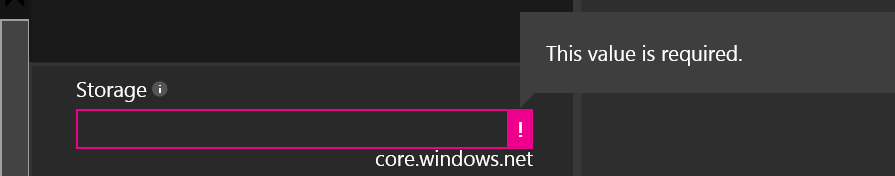


Figure 1‑9

Once the chosen name is valid, you’ll get a check green mark that tells you that the name is ok.

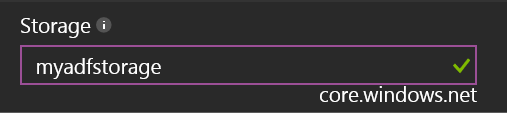


Figure 1‑10

##### Pricing tier

The pricing tier blade allows you to select the kind of storage you want to use. For the purpose of this book, we’ll use the basic (cheapest) one. But depending on your service level agreement (SLA) or requirements with your client or users, you might want to use a more expensive and performant storage.

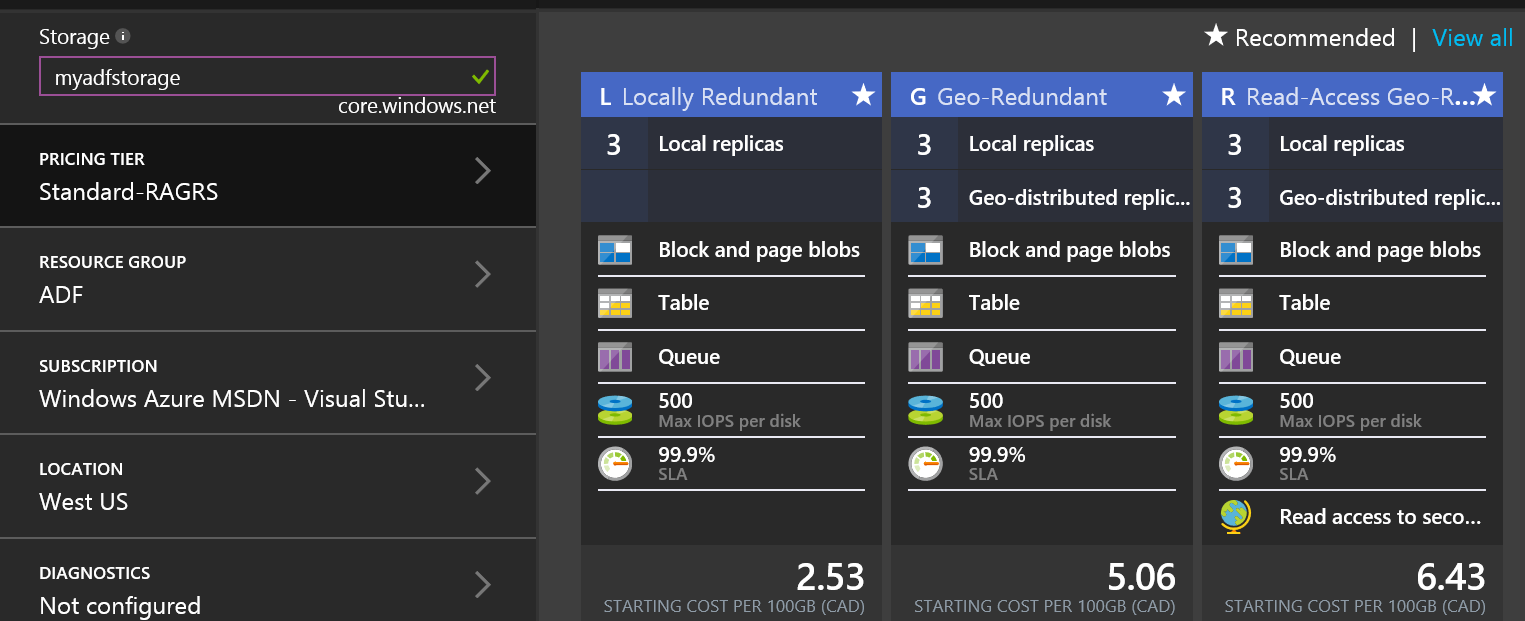


Figure 1‑11

If you click on “View all” as shown at the top right of Figure 1‑11, you’ll get more options in term of storage.

By default, all storage account types keep 3 copies of the files that you create or upload. This would allow file recovery if one file is damaged on a server for some reason. Some storage account options offer the possibility to geographically replicate and distribute your files at different data centers around the world. This gives more security and access to secondary storage via geographically distributed location.

With other storage options, you can have faster disk operations but at a higher cost. Prices varies upon time and are for each 100GB of monthly storage at time of writing.

##### Resource group

A resource group allows you to centralize resources into a common group in Azure. Think about them as folders in Windows/ A folder can have different types of files and / or folders. For example, let’s pretend you create a resource group named MyResourceGroup and you use it for factories, storage, HDInsight clusters and machines. When you decide to delete the resource group, all contained Azure resources contained into it will be deleted too.

You can access your resource groups in the Azure portal by clicking on “Browse” as shown on Figure 1‑12.

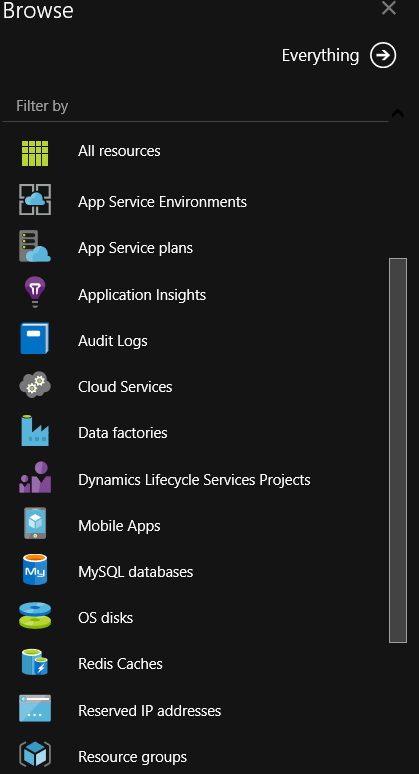


Figure 1‑12

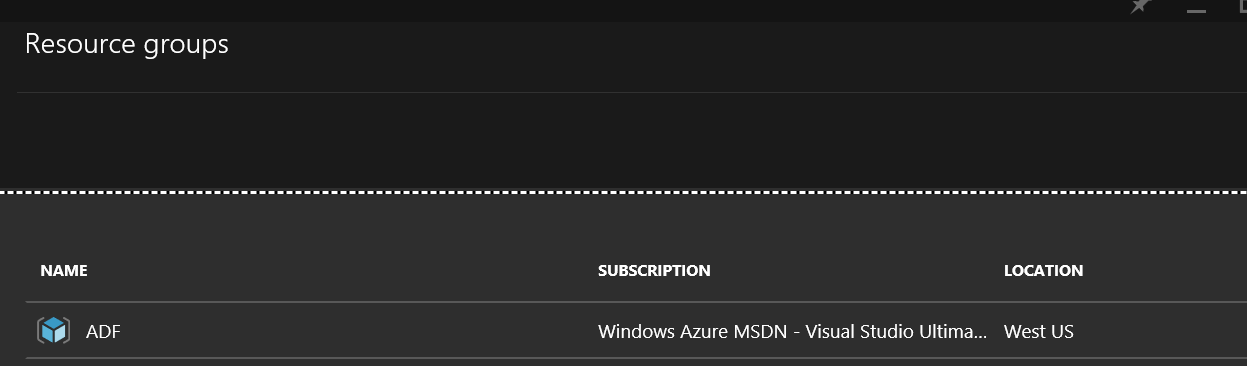
The Resource groups Azure Portal blade now appears.

Figure 1‑13

We can’t create resource groups individually, they have to be created when we create individual resources like storages, factories, etc. Every time theses individual resources are created, we have the option to specify a resource group.

##### Subscription

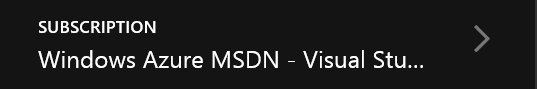


Figure 1‑14

The subscription blade allows you to select the Azure subscription you are using (and which will be invoiced) to interact with the storage. As we have talked about in Pricing tier section, depending on the type of storage you’ll be using and, as we’ll talk about in the Location section, storage is not free. You receive a monthly invoice for it as part of your Azure subscription.

#### Location

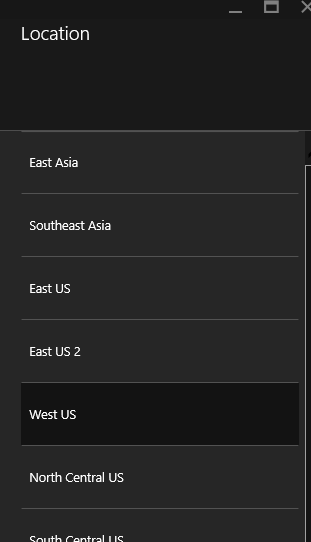


Figure 1‑15

When you create your storage resource, you have the option to create in different data centers in the world. The location that you choose has implications only if you plan to transfer data from one storage resource zone to another. And it depends if the data arrives in the zone (inbound) or goes out of it (outbound). For inbound traffic, there are no charges. But you will have to pay for outbound data or file movements.

Here are the various zones available at time of writing:

•Zone 1: US West, US East, US North Central, US South Central, US East 2, US Central, Europe West, Europe North

•Zone 2: Asia Pacific East, Asia Pacific Southeast, Japan East, Japan West, Australia East, Australia Southeast

•Zone 3: Brazil South

Notice that there are no charges for data transfer between locations that are in the same zone. For example, if you use Azure Data Factory to copy data from a blob storage that is located fin East-US to a SQL database located in Asia, depending on how much data we are transferring, charges may be applied. At time of writing, all monthly transfers below 5 gigabytes are free. You can refer to Microsoft Azure Data Transfers Pricing Details ( <http://azure.microsoft.com/en-us/pricing/details/data-transfers/#>) if you want to see all available options related to data transfers between zones.

## Sample case used throughout this book

The case we’ll use in this book is based on AdventureWorks LT database. This database was available on codeplex for all recent SQL Server edition up to SQL Server 2012. On the top of this database, we created a small data warehouse called AdventureWorksLTDW. This book’s support site will provide the necessary SQL Server data files for both databases.

### AdventureWorksLT database

This database contains bike accessories sales data. The database diagram (Figure 1‑16) describes the data model.

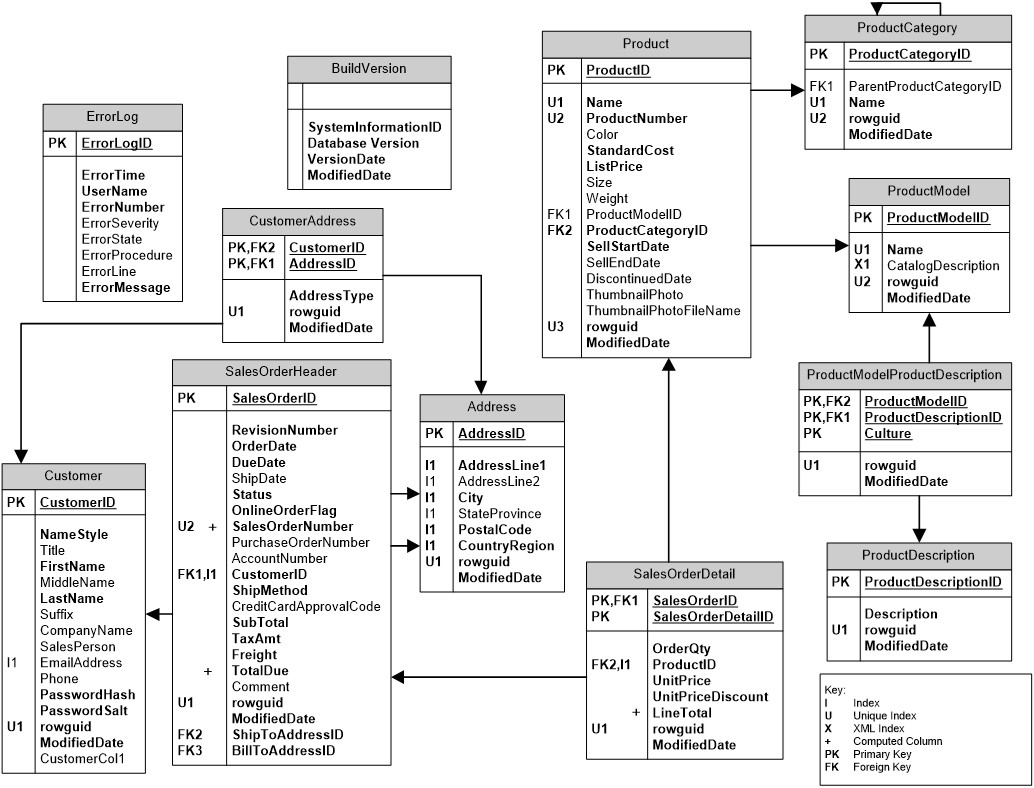


Figure 1‑16

### AdventureWorks LT DW

This database is a data warehouse that have been built on top of the