Movie recommendations on Azure

01/09/2019 • 4 minutes to read • Contributors

In this article

This example scenario shows how a business can use machine learning to automate product recommendations for their customers. An Azure Data Science Virtual Machine (DSVM) is used to train a model on Azure that recommends movies to users based on ratings that have been given to movies.

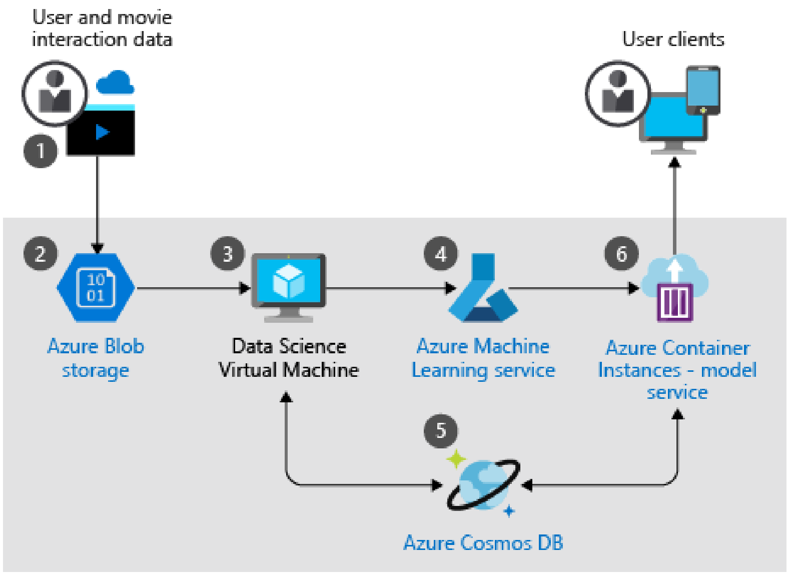
Recommendations can be useful in various industries from retail to news to media. Potential applications include providing product recommendations in a virtual store, providing news or post recommendations, or providing music recommendations. Traditionally, businesses had to hire and train assistants to make personalized recommendations to customers. Today, we can provide customized recommendations at scale by using Azure to train models to understand customer preferences.

Relevant use cases

Consider this scenario for the following use cases:

- Movie recommendations on a website.
- Consumer product recommendations in a mobile app.
- News recommendations on streaming media.

Architecture


User and movie interaction data
User clients
1
2
3
4
6
5
Azure Blob storage
Data Science Virtual Machine
Azure Machine Learning service
Azure Container Instances - model service
Azure Cosmos DB

This scenario covers the training and evaluating of the machine learning model using the Spark alternating least squares (ALS) algorithm on a dataset of movie ratings. The steps for this scenario are:

1. The front-end website or app service collects historical data of user-movie interactions, which are represented in a table of user, item, and numerical rating tuples.

2. The collected historical data is stored in a blob storage.
3. A DSVM is often used to experiment with or productize a Spark ALS recommender model. The ALS model is trained using a training dataset, which is produced from the overall dataset by applying the appropriate data splitting strategy. For example, the dataset can be split into sets randomly, chronologically, or stratified, depending on the business requirement. Similar to other machine learning tasks, a recommender is validated by using evaluation metrics (for example, precision@*k*, recall@*k*, MAP, nDCG@k).
4. Azure Machine Learning service is used for coordinating the experimentation, such as hyperparameter sweeping and model management.
5. A trained model is preserved on Azure Cosmos DB, which can then be applied for recommending the top *k* movies for a given user.
6. The model is then deployed onto a web or app service by using Azure Container Instances or Azure Kubernetes Service.

For an in-depth guide to building and scaling a recommender service, see Build a real-time recommendation API on Azure.

Components

- Data Science Virtual Machine (DSVM) is an Azure virtual machine with deep learning frameworks and tools for machine learning and data science. The DSVM has a standalone Spark environment that can be used to run ALS.
- Azure Blob storage stores the dataset for movie recommendations.
- Azure Machine Learning service is used to accelerate the building, managing, and deploying of machine learning models.
- Azure Cosmos DB enables globally distributed and multi-model database storage.
- Azure Container Instances is used to deploy the trained models to web or app services, optionally using Azure Kubernetes Service.

Alternatives

Azure Databricks is a managed Spark cluster where model training and evaluating is performed. You can set up a managed Spark environment in minutes, and autoscale up and down to help reduce the resources and costs associated with scaling clusters manually. Another resource-saving option is to configure inactive clusters to terminate automatically.

Considerations

Availability

Machine-learning-built apps are split into two resource components: resources for training, and resources for serving. Resources required for training generally do not need high availability, as live production requests do not directly hit these resources. Resources required for serving need to have high availability to serve customer requests.

For training, the DSVM is available in multiple regions around the globe and meets the service level agreement (SLA) for virtual machines. For serving, Azure Kubernetes Service provides a highly available infrastructure. Agent nodes also follow the SLA for virtual machines.

Scalability

If you have a large data size, you can scale your DSVM to shorten training time. You can scale a VM up or down by changing the VM size. Choose a memory size large enough to fit your dataset in-memory and a higher vCPU count in order to decrease the amount of time that training takes.

Security

This scenario can use Azure Active Directory to authenticate users for access to the DSVM, which contains your code, models, and (in-memory) data. Data is stored in Azure Storage prior to being loaded on a DSVM, where it is automatically encrypted using Storage Service Encryption. Permissions can be managed via Azure Active Directory authentication or role-based access control.

Deploy this scenario

Prerequisites: You must have an existing Azure account. If you don't have an Azure subscription, create a free account before you begin.

All the code for this scenario is available in the Microsoft Recommenders repository.

Follow these steps to run the ALS quickstart notebook:

1. Create a DSVM from the Azure portal.
2. Clone the repo in the Notebooks folder:

   shell Copy

   ```shell
   cd notebooks
   git clone https://github.com/Microsoft/Recommenders
   ```

3. Install the conda dependencies following the steps described in the SETUP.md file.
4. In a browser, go to your jupyterlab VM and navigate to `notebooks/00_quick_start/als_pyspark_movielens.ipynb`.
5. Execute the notebook.

Related resources

For an in-depth guide to building and scaling a recommender service, see Build a real-time recommendation API on Azure. For tutorials and examples of recommendation systems, see Microsoft Recommenders repository.