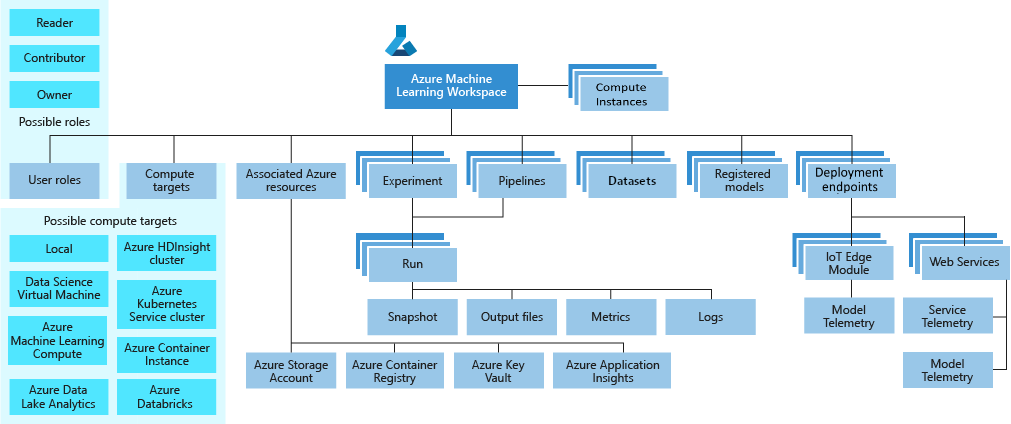
# Azure Machine Learning Services

Az Azure és a Machine Learning Services környezetben elérhető szolgáltatások, entitások térképe:



Az egyes fontossabb kifejezések magyarázata:

*Workspace:* Azure Machine Learning munkavégzési terület, ahol modelleket lehet futtatni/tanítani/összehasonlítani/tárolni, tanítási folyamatot nyomon követni, metrikákat számolni, naplózni, adatszetteket tárolni.

*Run*: Egy kódrészlet, amely valamilyen specifikus feladatot teljesít.

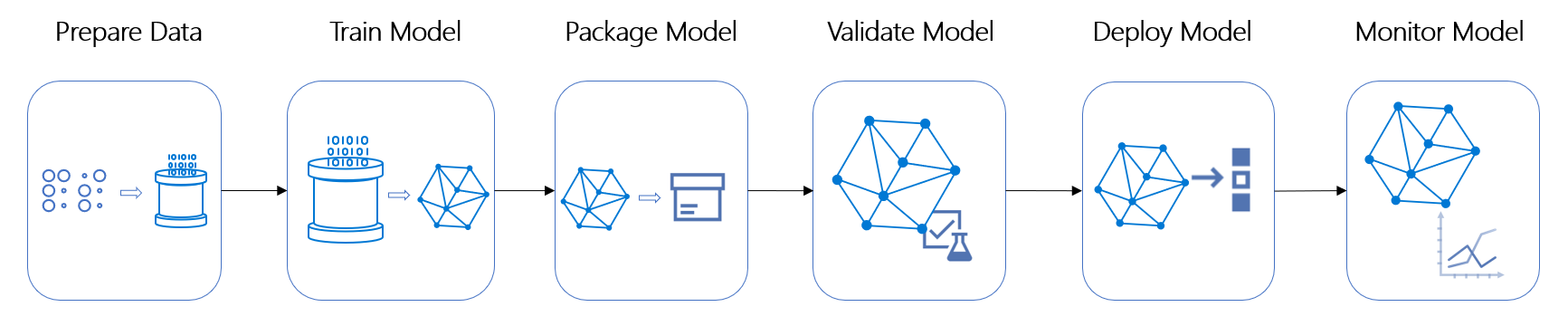
*Experiment*: Run-ok gyűjteménye.

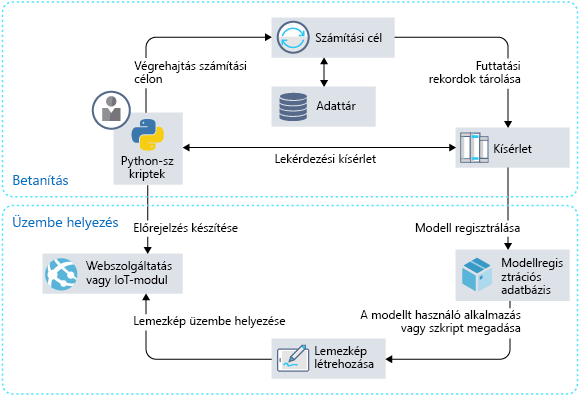
*Model:* tetszőleges betanított modellt el lehet tárolni a munkaterületen, amit a model regisztrálásával (register model) lehet elérni. A regisztrálandó modell lehet az AutoML funckiónak a terméke, vagy lehet egy másik könyvtárból exportált modell, például egy scikit-learn-ből kimentett Pickle vagy ONNX formátumú modell.

*Pipeline:* egy olyan tárolt munkafolyamat, ami több alfeladat szekvenciája. Egy pipeline lehet egyetlen Python szkript hívás, de általában egy komplexebb folyamat, magába foglalva az adatok előfeldolgozását, paraméterek inicializálása/beállítása,

futtató erőforrások konfigurálása, model betanítása, modell hyperparaméter optimalizálása , modell üzembe helyezése, stb… Egyetlen forrásba szervezett, monolitikus projekt helyett kézenfekvő a logikailag elkülünölő részeket szétválasztani külön forrásokba, majd ezeket egy Pipeline-ba szervezni.

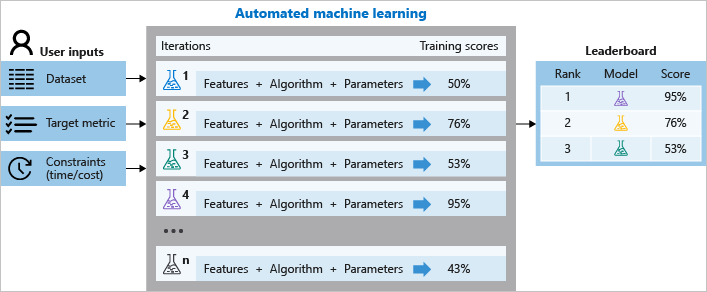
A Pipeline a függőségkezelése nagyon hatékony: minden változás esetén a rendszer megvizsgálja milyen szoftver/hardwer/operációs rendszer követelmények változtak, és elvégzi a szükséges változtatásokat, a legoptimálisabb megoldás megkeresésével.



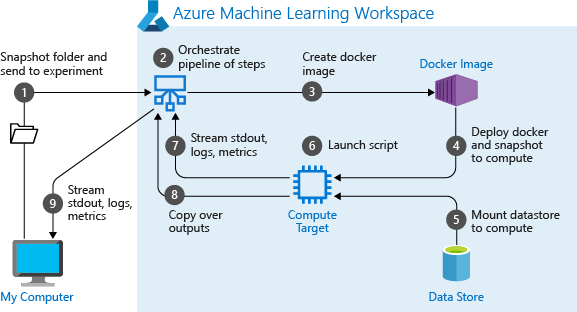


*HyperDrive:* Olyan szolgáltatás, amely a hiperparaméterek finomhangolását automatizálja, elindítva több különböző paraméterekkel konfigurált experiment-et, majd kiválasztva azt, amely az adott hiperparaméterek mellett a legjobban teljesít.

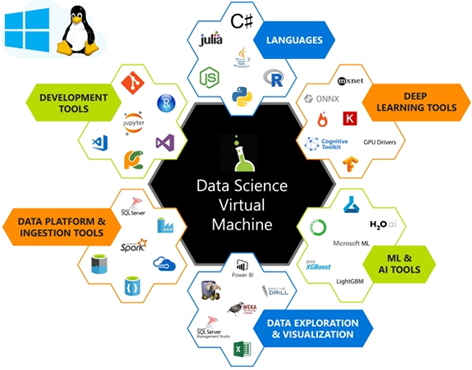
*AutoML:* Egy olyan szolgáltatás, amely egy adott feladatra **több** gépi tanulás modellt párhuzamosan lefuttat, megspórolva ezáltal a felhasználó modell kiválasztására fordítandó munkáját, idejét. Az algoritmus több fajta modellt, több fajta optimalizáló módszer, illetve több fajta előfeldolgozási (normalizációs) módszert figyelembe vesz, ezeknek minden kombinációját kipróbálja. Többek között, ensemble (aggregált, több modell együttese) modelleket is figyelembe vesz. A normalizáció alapértelmezetten része a folyamatnak, de hozzá lehet adni újabb funkciókat, például one-hot elkódolás, hiányzó cellák kezelése. Az felhasználható normalizációs algoritmusok listája: [link](https://docs.microsoft.com/en-us/azure/machine-learning/service/concept-automated-ml#preprocess). Az algoritmus továbbá regularizációt is alkalmaz a túltanulás elkerülése érdekében.



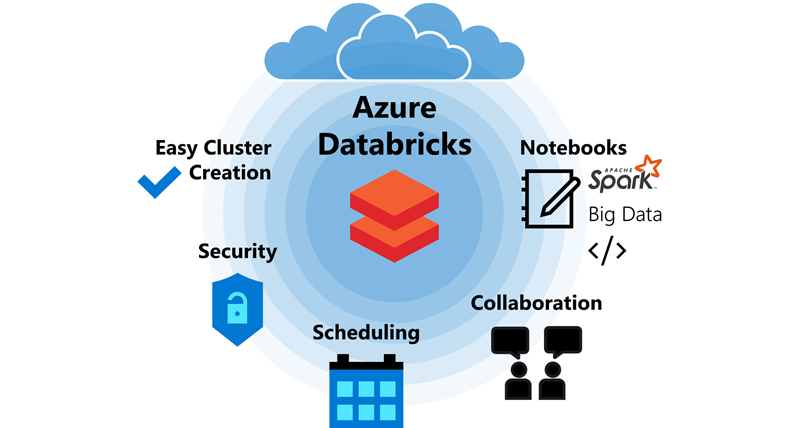
*Compute Target:* olyan erőforrás, ami a gépi tanulás megoldásokat lefuttatja, végrehajtja, például: virtuális gép, helyi gép, Azure Databricks, stb…



*Data Science Virtual Machine:* adattudományi könyvtárakkal előkonfigurált virtuális szerver (fizikai számítógép emulációja, amelyen egy operációs rendszer fut, amit távoli hozzáféréssel lehet használni, managelni). Ez egy IaaS (Infrastructure-as-a-Service) típusú szolgáltatás, azaz a szükséges hardware-t a rendszer gombnyomásra hozzárendeli a felhasználó munkaterületéhez, ezekkel nem kell a felhasználónak foglalkoznia. Virtuális gép esetében viszont a szoftvertkörnyezetet, az operációs rendszert a felhasználónak kell karban tartania, ehhez a hozzáférése is megvan természetesen. Lehetőség van egy, vagy akár több virtuális gép létrehozására is, Windows vagy Linux alapú operációs rendszerekkel.



*Azure Databricks:* egy összesített kollaboratív adattudományi platform, amelyhez Notebook interfész, feladat időzítő és egyéb Azure szolgálatásokkal való összeköttetés társul. A rendszer által támogatott nyelvek: Scala, Java, SQL, Python, és R. Grafikus felületén keresztül hasznosítható az Azure Spark HDInsight számítási klaszter szolgáltatás.

**

*Azure Data Lake Analytics:* egy big-data központú adatanalitikai platform, amely nagy méretű adattárak tárolására, feldolgozására és elemzésére kidolgozott szolgáltatás. Adattárolási lehetőségei az Azure Blob Storage megoldásainál fejlettebbek, például hierarchikus rendszert használ, amely a fájlműveletek sebességét lényegesen felgyorsítja.

*Azure Machine Learning Compute:* egy speciális *compute target*, egy számítógép klaszter, amely több CPU/GPU felhasználásával számításigényes tanítási folyamatokat végez.

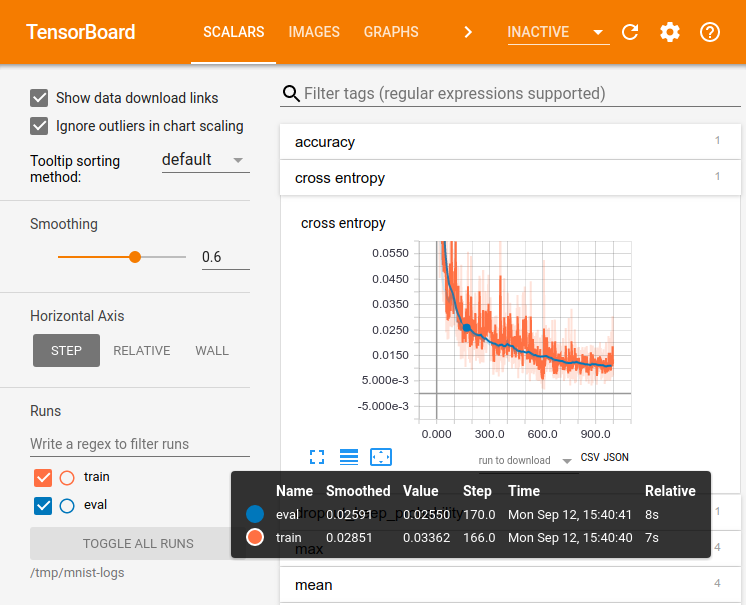
*Azure Container Instance:* konténer szolgáltatás (konténer: egy környezet, amely egy program és az általa szükséges könyvtárak, valamint futtatáshoz szükséges konfigurációs adatok szeparált gyűjteménye és az operációs rendszer fölött fut - előnyei a gyorsaság, egyszerűség és rendszerfüggetlenség a virtuális számítógéppel szemben)

*Azure Kubernetes Service:* nagy számú konténerek automatizálására, managelésére, kezelésére kifejlesztett szolgáltatás.

*Image:* a modell “képe”, amely tartalmazza magát a gépi tanulás modellt, egy programot, amely az input adatokat lefuttatja a modellen és visszaadja a kimenetet és a modell futtatásához szükséges könyvtárak gyűjteményét. Lehetőség van Docker kompatibilis image-t létrehozni.

*Deployment:* az Azure ML Services képes a rögzített image (modell-kép) közzétételére, mint web szolgáltatás vagy mint IoT modul.

*Tensorboard:* A Tensorflow által fejlesztett tanulófolyamat monitorizáló felületet az Azure is támogatja, és lehetőséget biztosít a felület felhasználására az ML Services SDK-n keresztül.



Bevezetés az Azure ML Service szolgáltatásba (lecke): [link](https://docs.microsoft.com/hu-hu/learn/modules/intro-to-azure-machine-learning-service/).

### Azure Machine Learning Service SDK

Lényegében egy SDK, mely a Machine Learning Service szolgáltatásainak hívását teszi lehetőve kódból. A helyi számítógépen való fejlesztést egészíti ki, felgyorsítja a betanítási folyamatot, segítséget nyújt a hiperparaméter finomhangolásban.

Néhány fontos osztálya/modulja:

* Workspace: workspace-ek kódból való manipulálását teszi lehetővé,
  + *create(...) -* létrehozás;
  + *get(...)* - meglévő workspace felhasználása
  + *get\_default\_datastore()* - a munkaterülethez társított adattára mutató változó lekérése
* Experiment:
  + konstruktor (*Experiment(ws, name)*)- experiment létrehozása egy adott munkaterületen, adott névvel.
  + *submit(config, ...)* -
* Run:
  + *Experiment.start\_logging()* -létrehoz egy run-t az experiment-en belül, amely elkezdi naplózni az adott futtatás metrikáit.
  + *run.get\_portal\_url()* - visszaad egy nyilvános url-t, amely az Azure Portálra irányít, ahol az adott run részletei, kimenete látható.
  + *register\_model()* - modell regisztrálása (mentése)
* Estimator
* AutoMLConfig: az AutoML automatikus modell-összehasonlító szolgáltatás konfigurációs osztálya.
  + konstruktor (*AutoMLConfig(...)*) - néhány fontos megadható paraméter: feladat típusa (klasszifikáció / regresszió), tanuló adatok, metrika (az optimalizálandó metrika), számítási erőforrás (compute target - milyen erőforráson fusson az algoritmus), feketelistás modellek (a folyamatból kihagyandó modellek listája), iterációk száma (minden iterációban kipróbál egy kombinációt az algoritmus), stb...
* ComputeTarget
* Dataset
* Model: gépi tanulás modelleket reprezentáló osztály
  + *register()* - modell regisztrálása lokális fájlból

### Azure CLI (Command Line Interface)

Az Azure CLI egy parancssor alapú program, amely az Azure erőforrások kezelésére használható. A CLI elérhető Windowsra, macOS-re és Linux rendszerekre. A CLI az Azure portálon való kezelést váltja ki, gyorsan kiadható szöveges parancsok formájában.

A parancssor lehetőséget biztosít erőforráscsoportok, tárolók, virtuális számítógépek, konténerek, gépi tanulás feladatok / folyamatok és egyebek adminisztrációjára. Bármilyen parancs kiadása előtt autentikáció szükséges, amellyel a felhasználó megadja, hogy melyik Azure előfizetést szeretné kezelni, majd bejelentkezik.

### Fejlesztői környezetek

* Azure Notebooks
* Visual Studio Code - Machine learning extension
* JetBrains PyCharm
* Jupyter Notebook

### 

### Azure Machine Learning Data Preparation SDK

Nagy számú előfeldolgozási eljáráshoz biztosít hozzáférést a fejlesztő számára. Használati esetei a Pandas könyvtáréhoz állnak a legközelebb.

Néhány fontos funkció:

* Új oszlop hozzáadása meglévő oszlopokból: reguláris kifejezés alapján, oszlopokon végzett műveletek eredményéből, karakterlánc részét kivágva, oszlopok duplikálása…
* Oszlopok törlése, átnevezése, adattípus konverziók
* Sorok szűrése feltétel, reguláris kifejezés alapján
* Hiányzó adatok kitöltése átlag/medián/min/max értékekkel
* Statisztikák kigyűjtése, vizualizálása
* Normalizálás (MinMax)
* One-hot és label elkódolások
* Aggregáló függvények: összeg, átlag, medián, leszámolás, szórás.

Github repó az SDK-hoz készített minta-notebookokkal: [link](https://github.com/microsoft/AMLDataPrepDocs/tree/master/how-to-guides). Az SDK dokumentációja: [link](https://docs.microsoft.com/en-us/python/api/azureml-dataprep/?view=azure-ml-py).

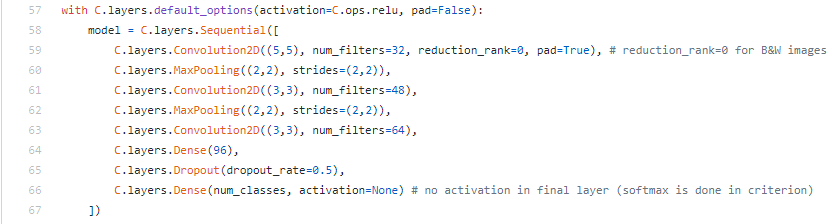
### Microsoft Cognitive Toolkit

A Microsoft Cognitive Toolkit (CNTK) egy deep learning keretrendszer, amellyel tetszőleges neurális hálókat lehet definiálni. Működése a Tensorflow könyvtárhoz hasonló, ahol a hálók szimbolikus számítási gráfok definíciójával jönnek létre. A Keras magas szintű mélytanulás könyvtár alapjául a CNTK-t is be lehet állítani. A CNTK összehasonlítása egyéb meglévő mélytanulásos könyvtárakkal: [link.](https://en.wikipedia.org/wiki/Comparison_of_deep-learning_software)

A CNTK könyvtárhoz C++, C# és Python programozási nyelvekben készültek API-k, illetve van egy saját neurális háló modell leíró nyelve, a BrainScript.

Támogatja a Open Neural Network Exchange (ONNX) formátumot, ami egy olyan eszköz, amivel betanított modelleket lehet egyik platformról a másikra exportálni, egyik könyvtárban megírt modellt felhasználni egy másik könyvtárban.

Az iparban használatos összes neurális háló típus felhasználására lehetőség van: hagyományos modellek, teljesen összekötött hálók, konvolúciós hálók ([mintafeladat](https://cntk.ai/pythondocs/CNTK_103D_MNIST_ConvolutionalNeuralNetwork.html)), rekurenns hálók, autoenkóder, generatív modellek, …

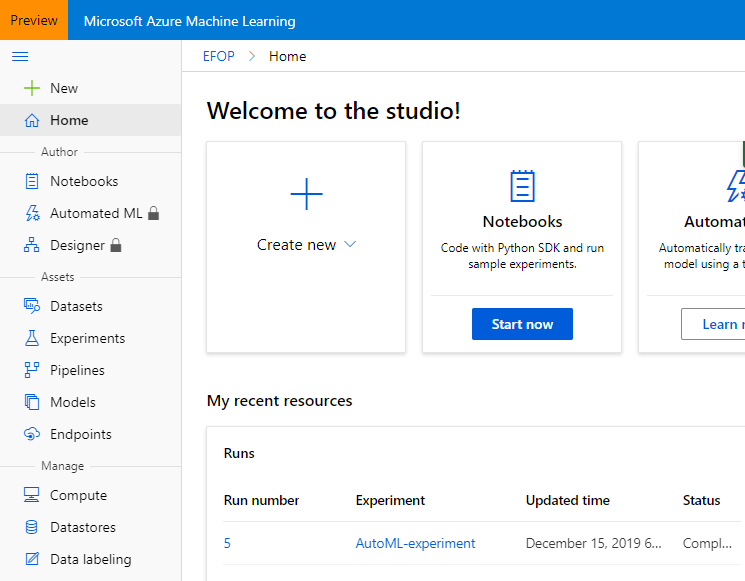


A mellékelt kódrészlet egy konvolúciós háló modell definiálását tartalmazza, látható, hogy egy magas szintű könyvtárról van szó, amely sok hasonlóságot mutat a Keras könyvtárral.

### Az új Azure Machine Learning Portal (ml.azure.com)

Nemrég a Microsoft megújította a Machine Learning Services-t, átemelve a Studio funkcionalitásait (új neve: Azure Machine Learning designer) a Services-be, létrehozva egy új felületet és kibővítve az eszköztárat új funkciókkal. Itt már lehetőség van saját kezűleg kitalált és betanított modell felhasználásra a designerben. A ML designer és a klasszikus ML Studio összehasonlítása: [link.](https://docs.microsoft.com/en-us/azure/machine-learning/compare-azure-ml-to-studio-classic#comparison-azure-machine-learning-vs-ml-studio-classic)

Az új funkciók egyike például a képek felcímkézésére kitalált dedikált felület (*data labeling* fül alatt).

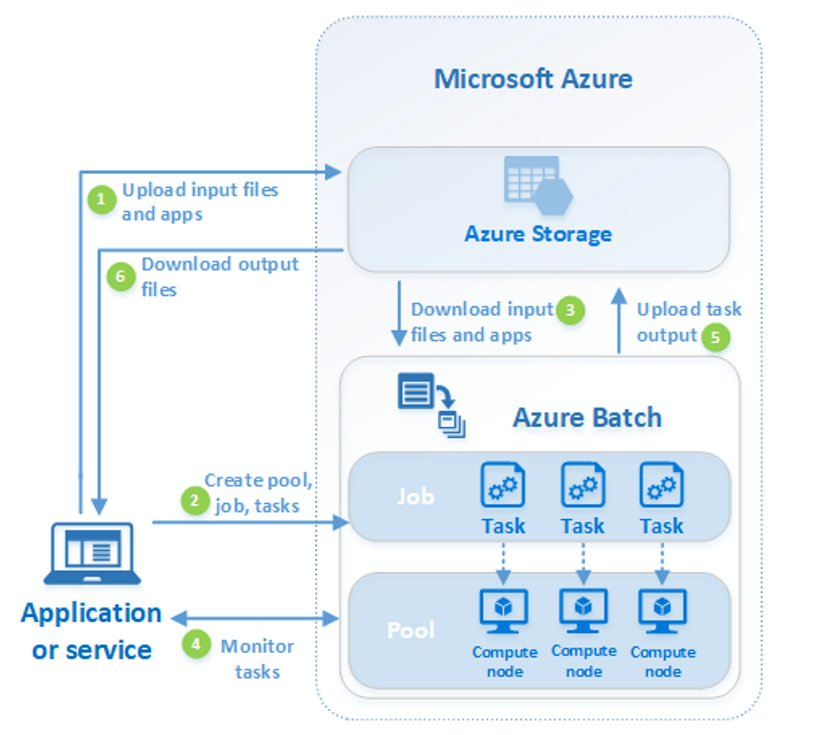


### Microsoft Azure Batch

Azure Batch egy nagyméretű feladat időzítő és managelő szolgáltatás, amellyel akár több ezer virtuális gép igénybevételét lehet vezérelni.

Amikor készen állunk egy komplex feladat lefuttatására, a Batch az alábbiakat végzi el:

* Elindít egy virtuális gép klasztert
* Feltelepíti a szükséges szoftvereket és kialakítja a környezetet
* Minden taszkot elindít, lefuttat
* Felismeri, naplózza a hibákat
* Az új, megismétlendő taszkokat sorba állítja, újraindítja
* Leállítja, lecsatlakoztatja az infrastruktúrát a feladat végén



### Microsoft Azure Batch AI

A Batch AI egy, már nem támogatott szolgáltatás, amely nagy számításigényű gépi tanulás feladatok párhuzamos futtatását és követését tette lehetőve. A Batch AI által nyújtott szolgáltatásokat az Azure Machine Learning Studió és ezen belül a Machine Learning Compute service vette át.