Azure for Data Science

Ivan Putera Masli
Kshitij Kumar

Data Science?

- Extraction of **previously unknown** and **potentially relevant** knowledge and insight from data
- Intersection of Computer Science, Mathematics, Business and Programming (sometimes hacking).

Why Data Science now?

1. Data is cheap and abundant
2. Development of new algorithms
3. Rise of cloud services

Azure

- Microsoft’s Cloud Computing solutions
- 50 regions served worldwide (as of May 2018)
- Second biggest cloud provider by market share

Data Science with Azure

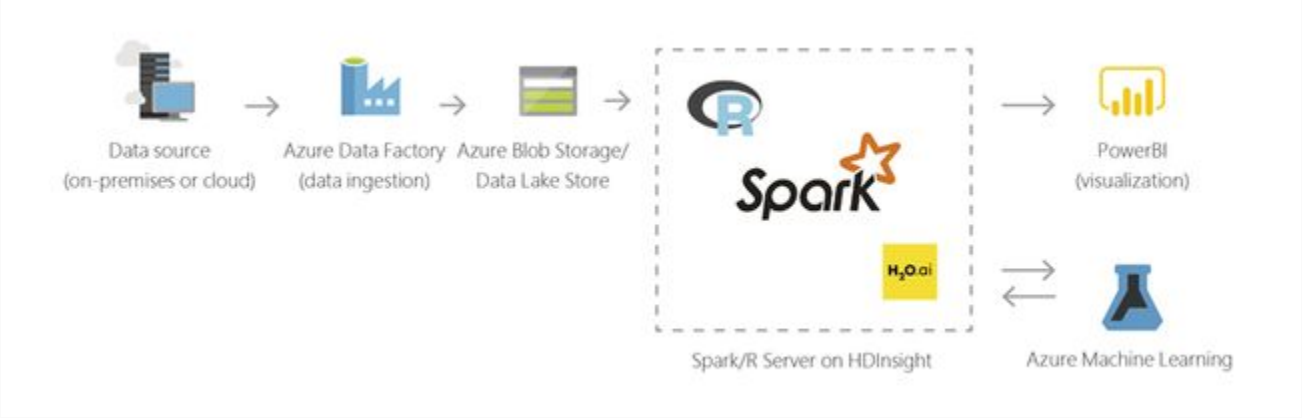
- More than a hundred services provided by Azure
- Example of services suitable for Data Science:
  - HDInsight
  - Machine Learning
  - AI
  - Data Science Virtual Machine

Azure HDInsight

- Cloud based solutions for Big Data Ecosystem (Hadoop, Spark, Kafka, etc)
- Fully managed platforms (Infrastructure, Scaling, Security)
- Less resource required to your Data Science platform maintenance
- 99,99% SLA (Less than 1 hour downtime each year)

Azure HDInsight for Data Science

- Fully Integrated with other Azure services



Azure HDInsight Customer Story

- PROS, a SaaS company which focussed on price optimization and revenue management.
- Allows functionality **specific** to running large **computations** with **huge amounts of data**
- Easily configured and run regardless of the size and the time required also shrank significantly.

Azure AI

Pre-built APIs: Cognitive Service APIs, ChatBot Services

Custom models with Azure Machine Learning, including Deep Learning models

- Azure Machine Learning Studio
- Azure Machine Learning Services
- Data Science Virtual Machine

Availability of a large number of open source *frameworks*: Tensorflow, MXNet, Chainer, PyTorch, Caffe, scikit-learn, CNTK

Frameworks

TensorFlow

Open source software library for high performance numerical computation.

Azure Cognitive Toolkit

Free, open-source, commercial-grade toolkit to train deep learning algorithms optimized for speech.

Pytorch

Scientific computing framework that puts GPUs first.

scikit-learn

Simple and efficient tools for data mining and data analysis

Onnx

An open format to represent deep learning models.

Caffe2

Lightweight, modular, and scalable deep learning framework.

MxNet

A flexible and efficient library for deep learning.

Chainer

A powerful, flexible, and intuitive framework for neural networks.

Cognitive Service APIs

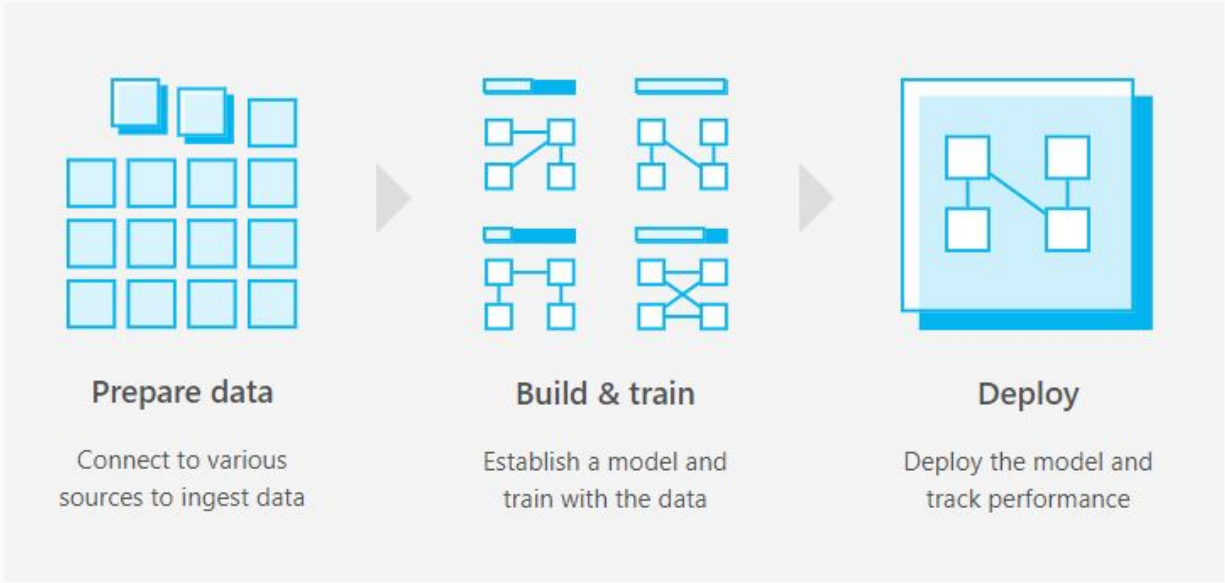
- Vision
  - Images, Faces, Video
- Language
  - Translation, Content moderation, Language understanding
- Speech
  - Speech <-> Text, Speaker recognition, Translation

Cognitive Service APIs

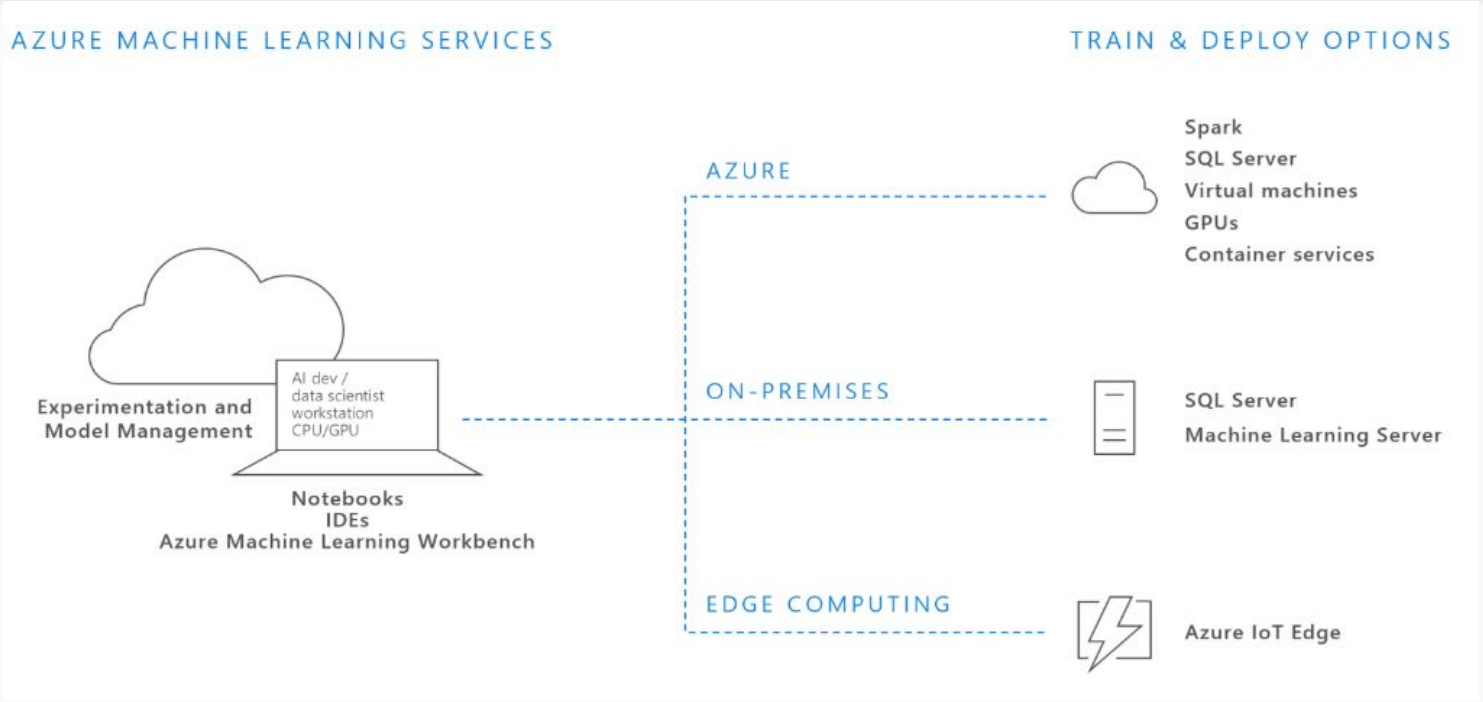
- Knowledge
  - Question Answer, Knowledge Bases)
- Search
  - using Bing: Visual, Video, News

Also, ChatBot services exist

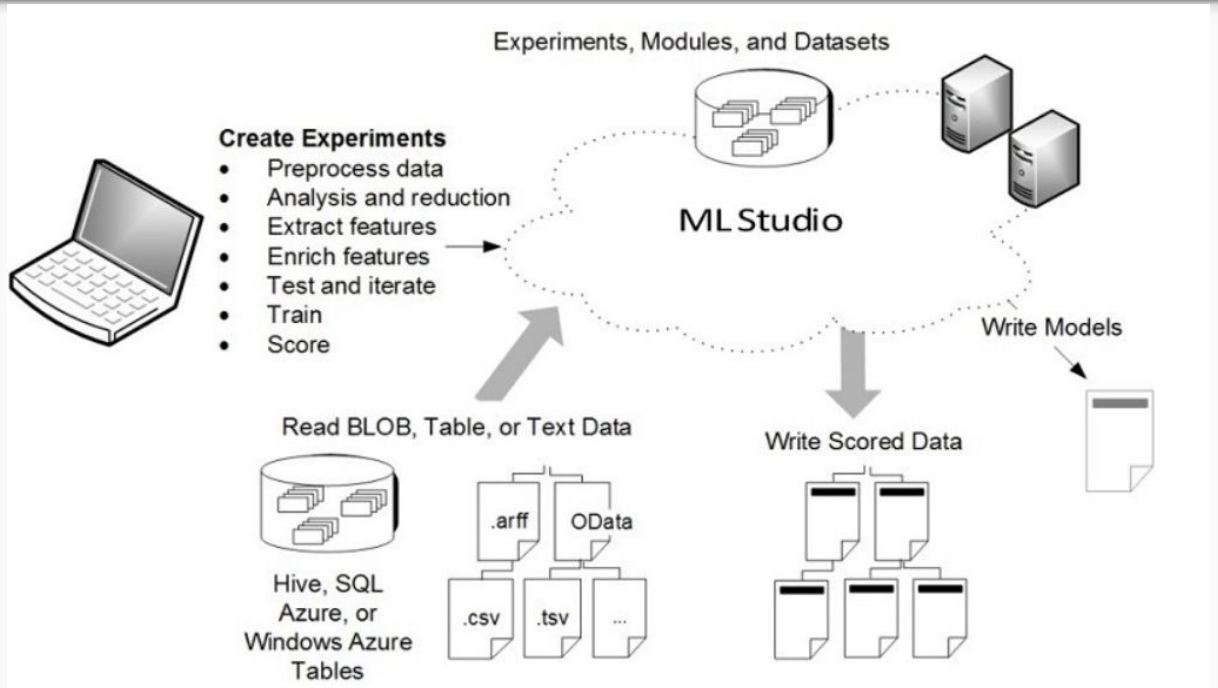
Data Science Pipeline



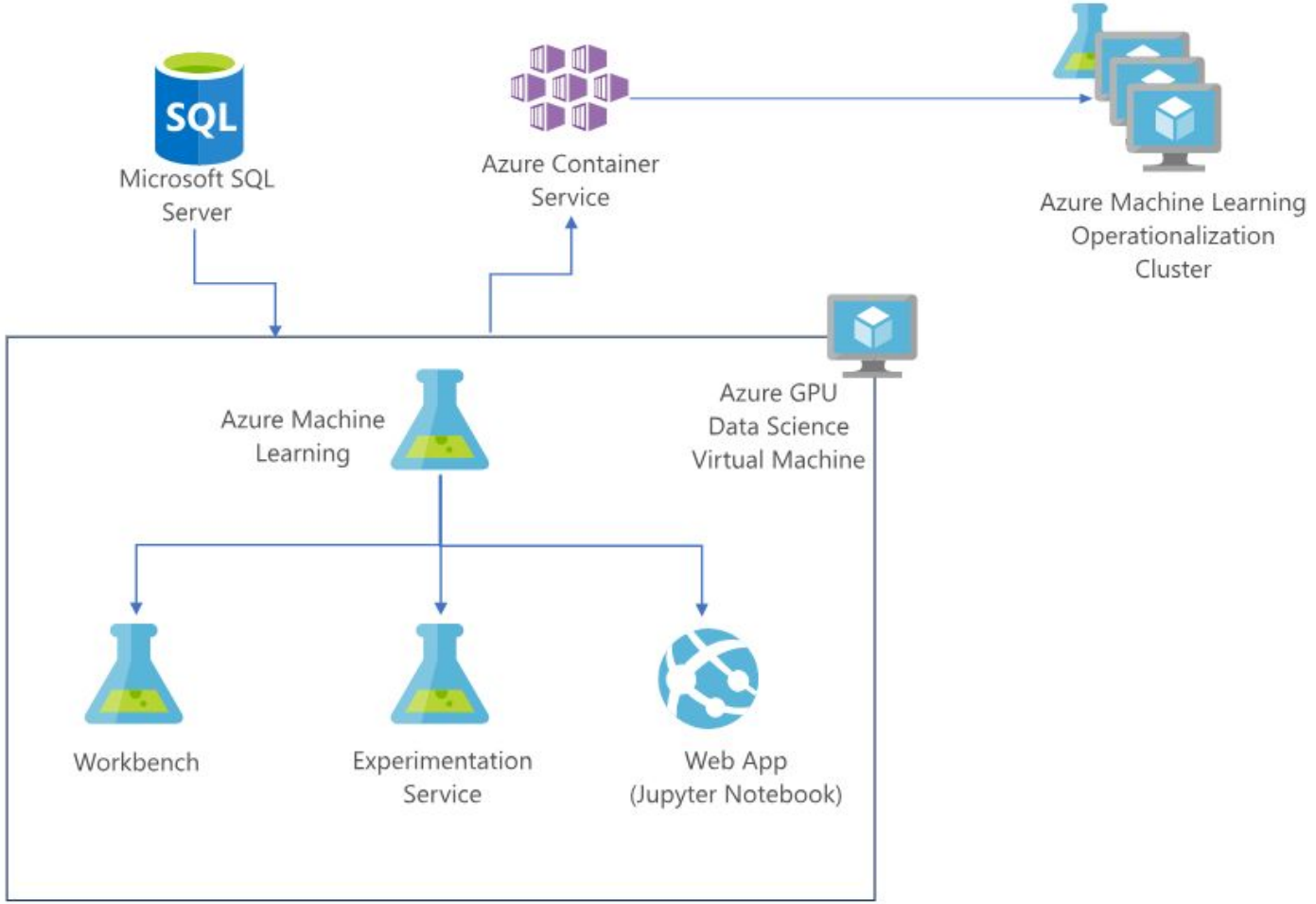
Azure ML Services Pipeline

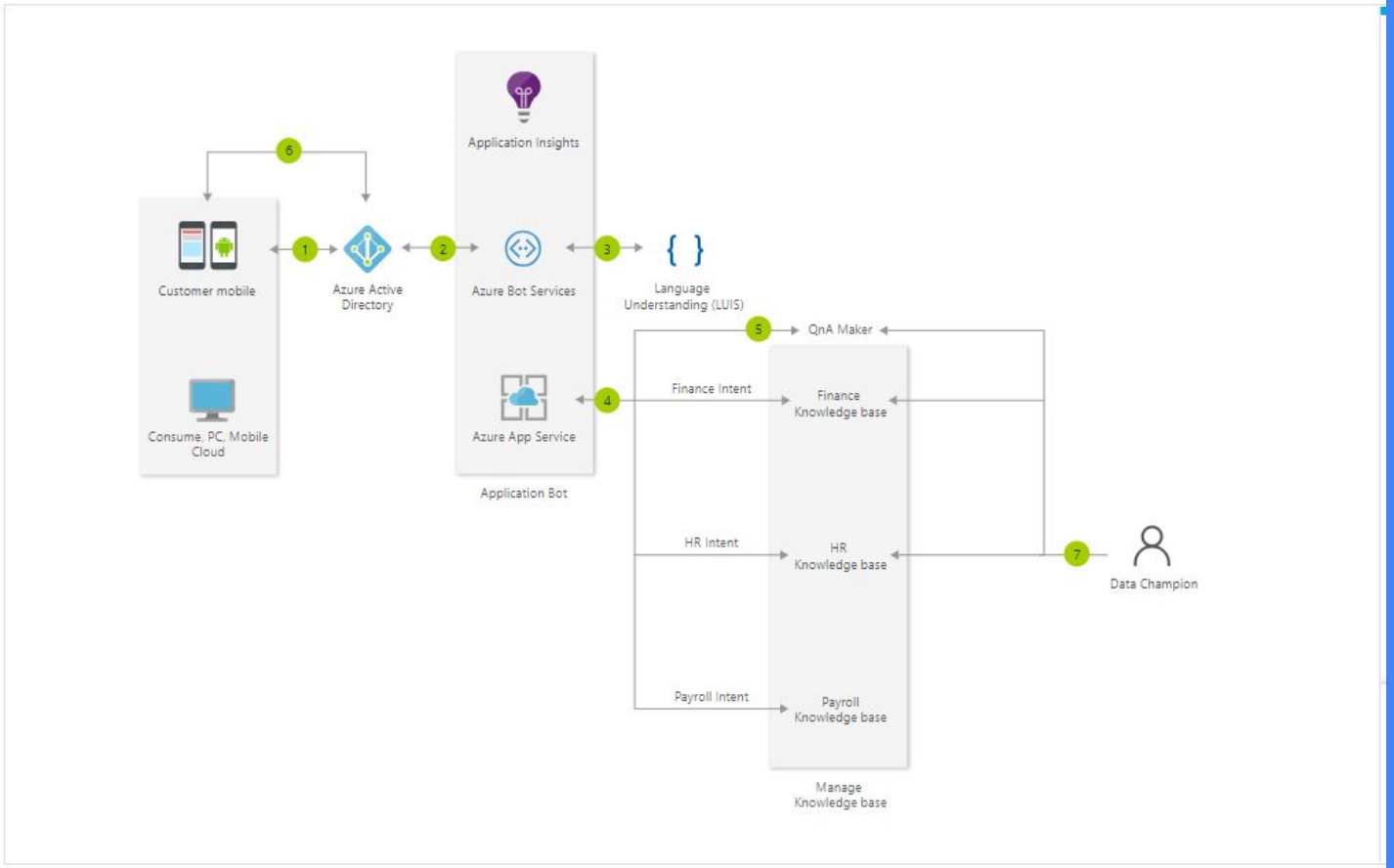


Azure ML Studio



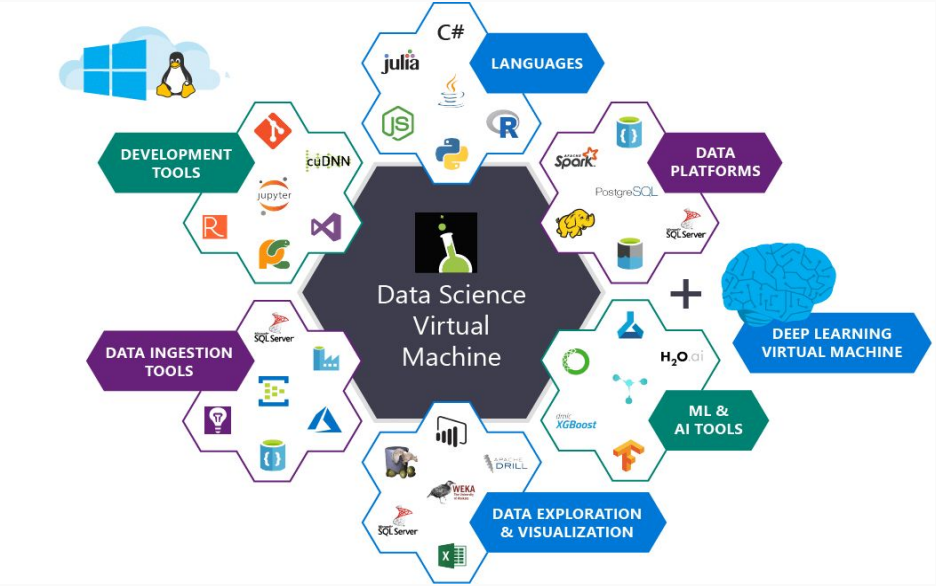
Case Studies: Azure API


SQL
Microsoft SQL Server
Azure Container Service
Azure Machine Learning Operationalization Cluster
Azure Machine Learning
Azure GPU Data Science Virtual Machine
Workbench
Experimentation Service
Web App (Jupyter Notebook)


Application Insights
6
Customer mobile
1
Azure Active Directory
2
Azure Bot Services
3
Language Understanding (LUIS)
5
QnA Maker
Consume, PC, Mobile Cloud
Azure App Service
4
Finance Intent
Finance Knowledge base
Application Bot
HR Intent
HR Knowledge base
7
Data Champion
Payroll Intent
Payroll Knowledge base
Manage Knowledge base

Data Science Virtual Machine (DSVM)

- Virtual Machine Image for Azure Virtual Machines
- Included various ready to use Software Stack, Framework and Platform for Data Science project
- **Key Advantage:** Saving your time from installing and configuring your favourite technology
- Suitable for fast prototyping, modeling and education



Azure Notebooks

- Notebook: A new way for doing (Data) Science
- Supports most commonly used Notebook (Jupyter)
- Supports Python 2, Python 3, F# and R
- Supports various plotting libraries such as ggplot, matplotlib, bokeh, and seaborn.
- Easily hosted and shared

PowerBI

- Business Analytics suite for Visualization
- Easily connected to wide range of data source
- Integrate aggregated to external information (e.g. Bing Map)
- Shareable dashboard

Thanks.

https://github.com/IllSc/CLOUD-COMPUTING-CLASS-2018/tree/master/Research-topic