Advance Data Mining
Lecture 11
23-Feb-2018

Special session on Cloud Technologies by

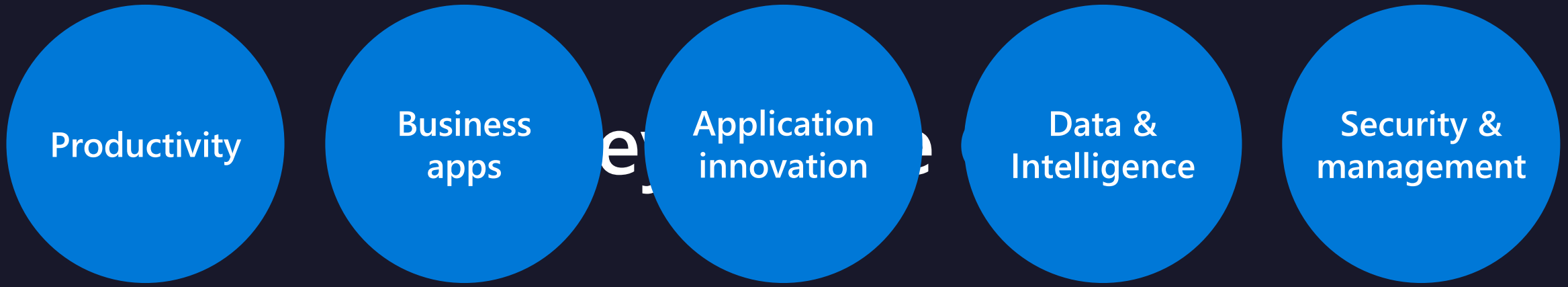
- Prof Kamlesh Tiwari, BITS Pilani
- Sudhir Rawat, Microsoft

Microsoft

Agenda

1. What is Cloud Platform
2. Introduction to Azure
3. Understanding of IAAS, PAAS and SAAS
4. Cloud, Application, Data Infrastructure
5. Processing Big Data on Azure
6. AI, ML and Data Science VM
7. Demo

What is the Cloud?


Productivity
Business apps
Application innovation
Data & Intelligence
Security & management
Office 365
Dynamics 365
Azure
Cortana Intelligence
Enterprise Mobility + Security
Operations Mgmt. + Security

Microsoft Cloud

Microsoft Azure

TOOLS

Visual Studio + Visual Studio Code + Visual Studio Team Services

ADVANCED WORKLOADS

Web + Mobile

Internet of Things

Microservices

Serverless

Identity Management

Data + Analytics

Cognitive Services

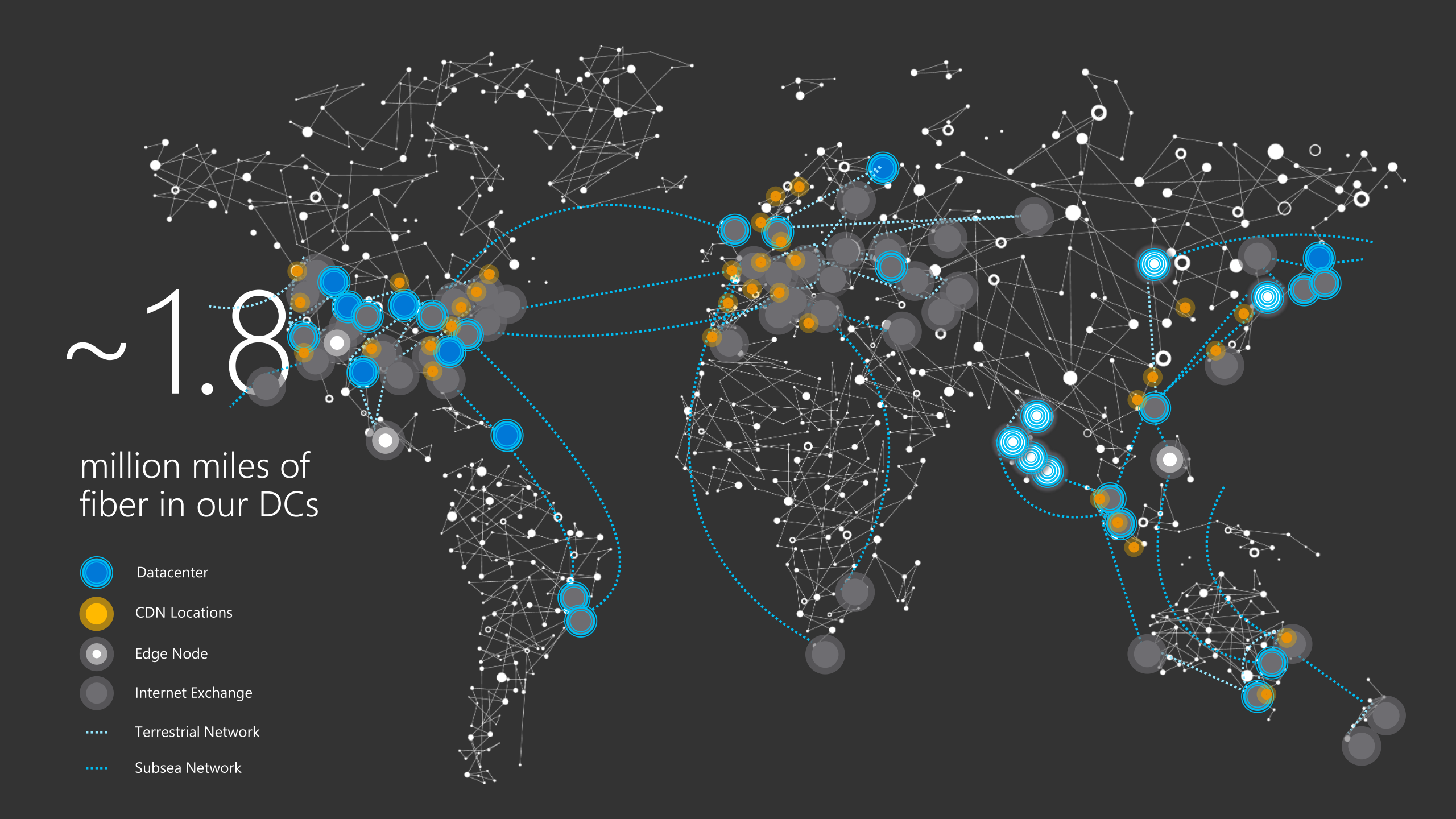
High Performance Compute

CORE INFRASTRUCTURE

Compute Storage Networking Security

Azure Stack + Hybrid

40 Azure regions


~1.8
million miles of
fiber in our DCs
Datacenter
CDN Locations
Edge Node
Internet Exchange
Terrestrial Network
Subsea Network

Azure: The Trusted Cloud

More certifications than any other cloud provider

GLOBAL

ISO 27001

ISO 27018

ISO 27017

ISO 22301

SOC 1 Type 2

SOC 2 Type 2

SOC 3

CSA STAR
Self-Assessment

CSA STAR
Certification

CSA STAR
Attestation

US GOV

Moderate
JAB P-ATO

High
JAB P-ATO

DoD DISA
SRG Level 2

DoD DISA
SRG Level 4

DoD DISA
SRG Level 5

SP 800-171

FIPS 140-2

Section 508
VPAT

ITAR

CJIS

IRS 1075

INDUSTRY

PCI DSS
Level 1

CDSA

MPAA

FACT UK

Shared
Assessments

FISC Japan

HIPAA /
HITECH Act

HITRUST

GxP
21 CFR Part 11

MARS-E

IG Toolkit UK

FERPA

GLBA

FFIEC

REGIONAL

Argentina
PDPA

EU
Model Clauses

UK
G-Cloud

China
DJCP

China
GB 18030

China
TRUCS

Singapore
MTCS

Australia
IRAP/CCSL

New Zealand
GCIO

Japan My
Number Act

ENISA
IAF

Japan CS
Mark Gold

Spain
ENS

Spain
DPA

India
MeitY

Canada
Privacy Laws

Privacy
Shield

Germany IT
Grundschutz
workbook

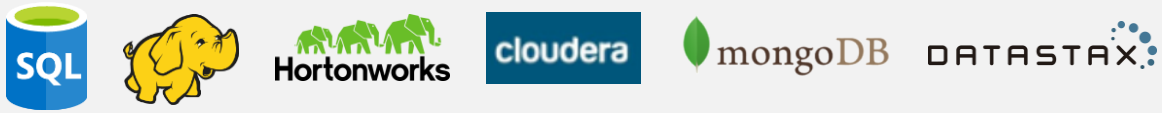
Choice

Management

Applications

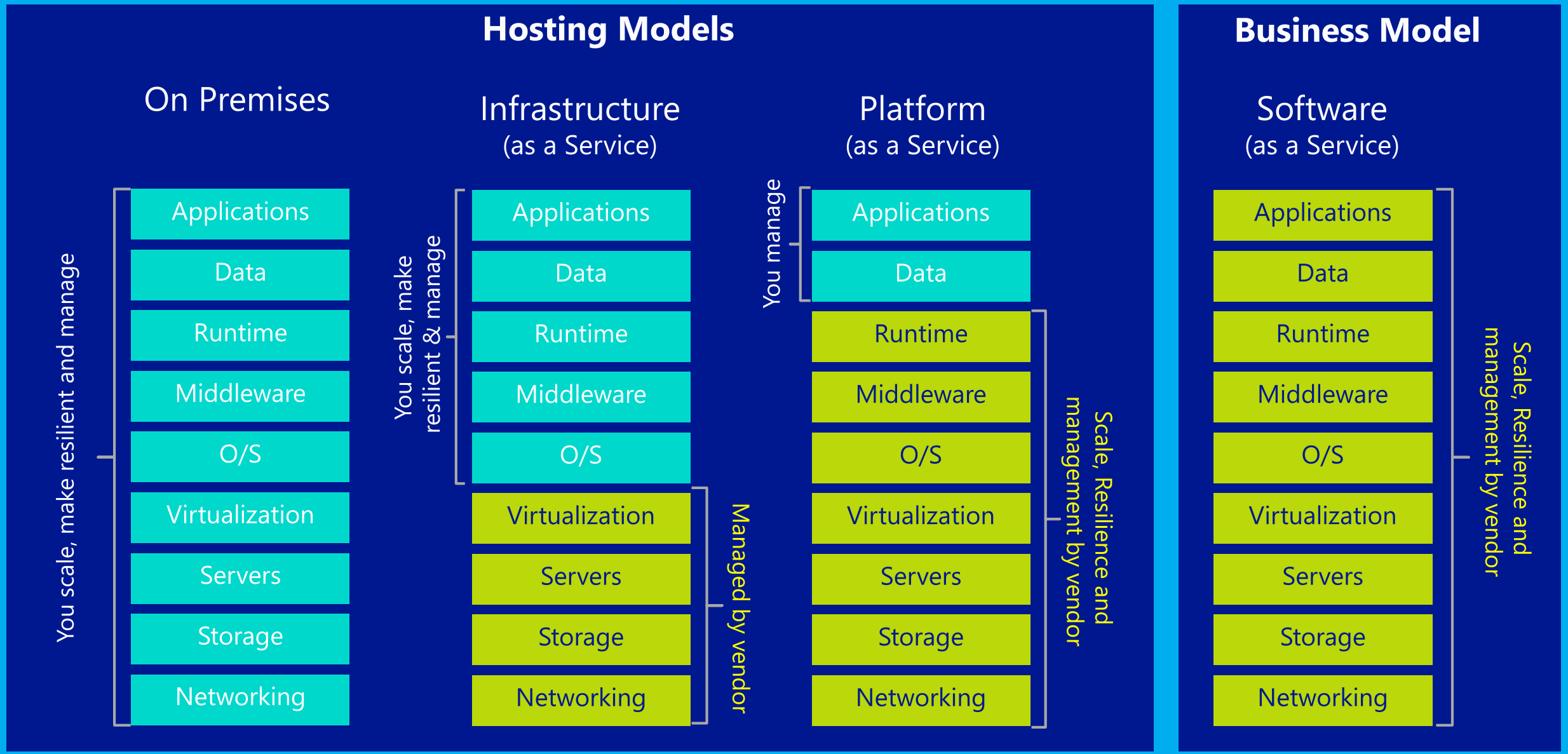
App Frameworks

Databases & Middleware



Infrastructure

Hosting & Cloud Software Delivery



Cloud Infrastructure

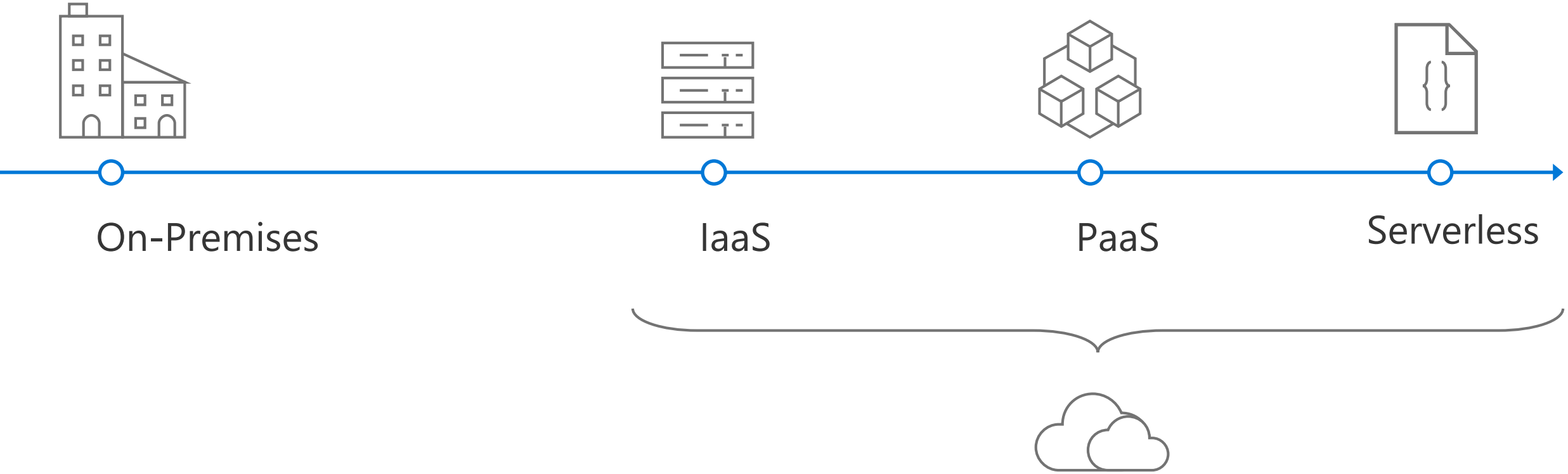
Virtual-Machine Sizes

Azure offers a variety of VM sizes and attributes

Series	Sizes	Attributes
A-Series	A0, A1, A2, A3, A5, A6, A7, A8, A9, A10, A11	Up to 112 GB RAM and 16 cores Up to 16 Data Disks (1 TB each)
D/DS/Dv2-Series	D1, D2, D3, D4, D11, D12, D14, D5v2	Up to 100% faster than A-Series Up to 112 GB RAM and 16 Cores Up to 32 data disks (1 TB each) Solid-state drives
G/GS-Series	G1, G2, G3, G4, G5	35% faster than D-Series Up to 448 GB RAM and 32 cores Up to 64 data disks (1 TB each) Solid-state drives

Application infrastructure

The “evolution” of application platforms



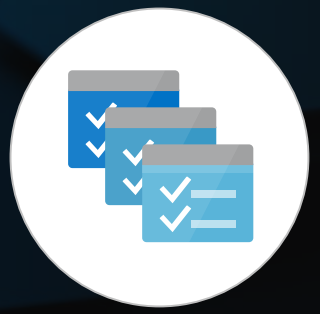
Deploy containers everywhere in Azure

Container Service

- **Kubernetes**
- **Mesos DC/OS**
- **Docker Swarm**

Service Fabric

Web Apps



Batch

Serverless application platform components

Development

- IDE support
- Visual debug history
- Local development
- Verbose debugging

Platform

Functions

- Developer tooling
- Bindings and triggers
- Open source

Logic apps

- Visual designer
- 100+ connectors
- Functions orchestration

Data/storage

Messaging

Gateway Connectors

Intelligence

Bots

Test

Analytics

Build

Distribute

Live Update

Crash Reporting

A/B Testing

Storage

Identity

Push Notifications

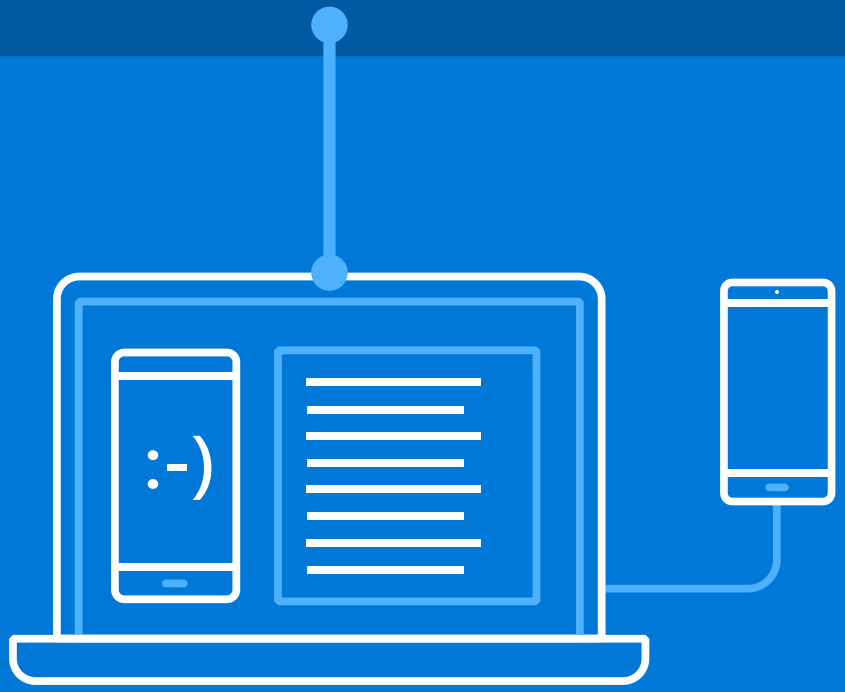
Remote Config.

Tables

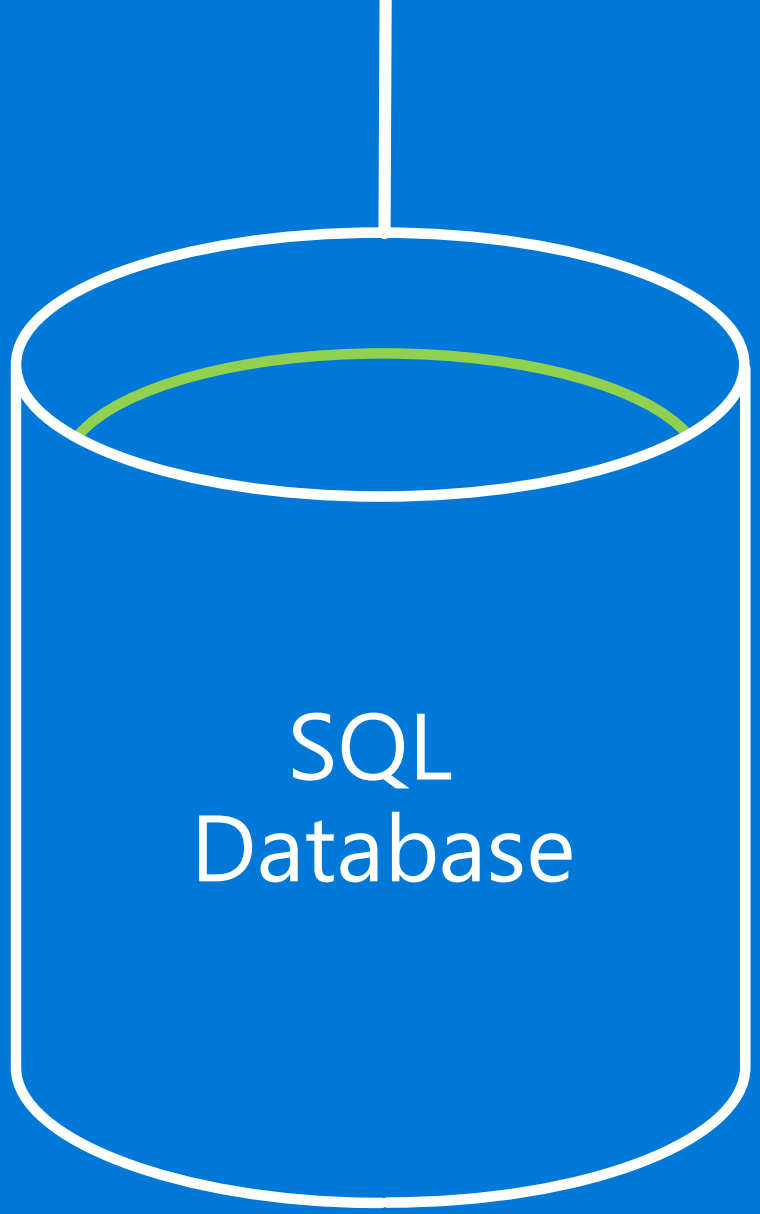
Visual Studio Mobile Center

Mission control for mobile apps
Windows, iOS and Andriod

Build
Test
Analytics
Push Notifications
Distribute
A/B Testing
Crash Reporting
Remote Config.
Identity
Storage
Live Update
Tables



Data infrastructure



Azure SQL Database

SQL Server as a Service

Highly available, durable, secure, fault tolerant—built-in

Scales without downtime, globally

In-memory for 30X performance improvements

Intelligence built-in for auto-tuning and threat detection

PostgreSQL as a Service
MySQL as a Service

Built-in high-availability and security

Elastically scale up or down with no application downtime

100% compatible with all existing drivers, libraries, tools

Azure Cosmos DB

The first globally distributed, multi-model database service

Globally distributed

Multi model, multi API

Scale to any need

Comprehensive SLAs

Azure Cosmos DB

Multi-model and multi-API

DATA MODEL	APIs
Key-value	DocumentDB
Document	MongoDB
Column Family	Table storage
Graph	Gremlin Graph
	Spark

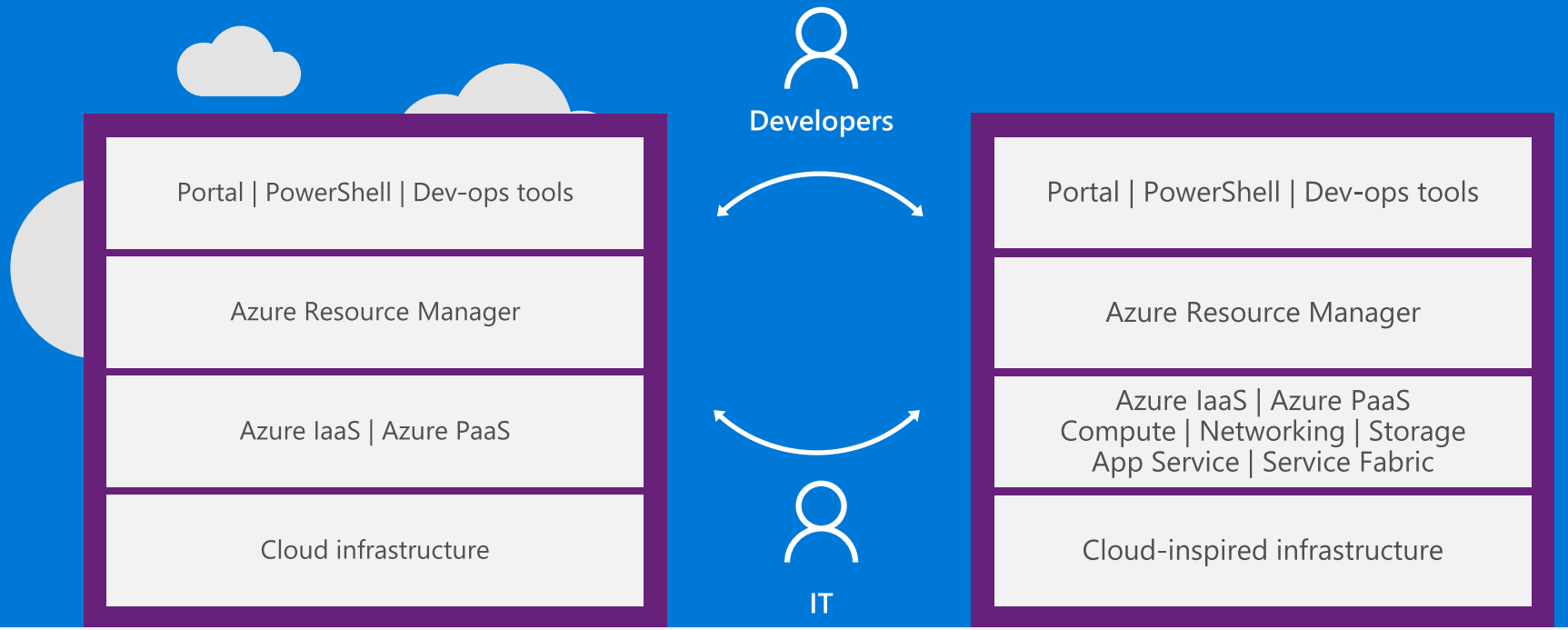
IoT Infrastructure

Azure IoT Suite

- Device Connectivity & Management
- Data Ingestion and Command & Control
- Stream Processing & Predictive Analytics
- Workflow Automation and Integration
- Dashboards and Visualization
- Initial SaaS for IoT approach (IoT Central)
- Solution Platforms
  - Connected Factory
  - Connected Vehicle & Mobility

Hybrid cloud platform

Azure Services in your datacenter



- Azure services in your datacenter
- Unified app development
- One Azure ecosystem

Microsoft Azure Public

Microsoft Azure Stack
Private | Hosted

Hadoop (HDInsight)

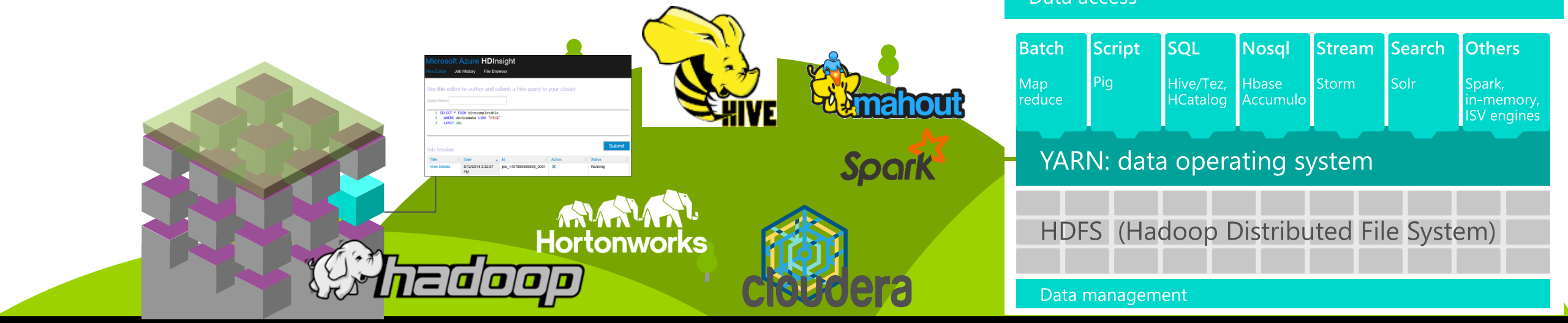
What is the feature?

Azure HDInsight is Microsoft's cloud Hadoop as a Service
- 100% open source Apache Hadoop
- Built on the latest releases across Hadoop (2.4)
- Up and running in minutes with no hardware to deploy (PaaS)
- Fully managed, operated and supported by Microsoft
- Utilize familiar BI tools for analysis including Microsoft Excel

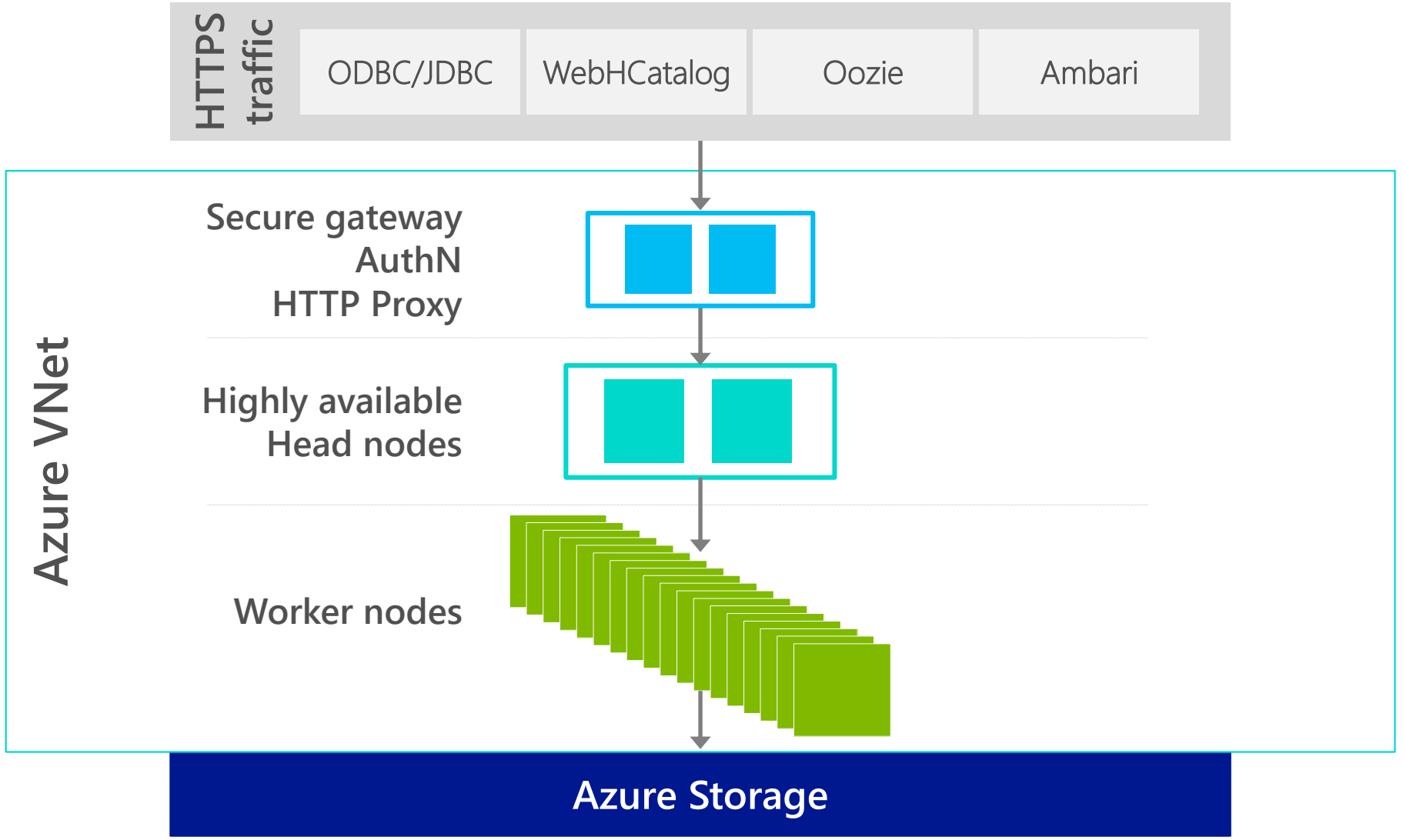
Comprises **core services** of MapReduce, HDFS, and YARN
Data services (Hive, HBase, Pig, Flume, Sqoop)
Operational services to manage the cluster (Ambari, Falcon, and Oozie)
- ODBC/JDBC connections to Hive
- Phoenix JDBC connectivity to Hbase
- REST control endpoints:
- Templeton – Job submission and management
- Ambari – Cluster monitoring
- YARN – YARN application submission
- Oozie – Orchestration and scheduling control

SPARK in memory 100x; framework only

	Hadoop in IaaS	HDP & CDH Gallery	HDInsight
OS Options	Any	CentOS	Windows, Ubuntu
Hadoop distribution	Any	Either HDP or CDH	HDP
Deployment	Up to user	From Azure Portal / Scripts	From Azure Portal / Scripts
Management	Up to user	Up to user	On your behalf
Support	Through vendor	Through vendor	Microsoft Azure
SLA (for Hadoop)	None	None	99.9%
OS Updates	Up to user	Up to user	Provided
Hadoop updates	Up to user	Up to user	Provided
Cluster Scaling	Up to user	No	Yes
Pricing	VM cost	VM Cost + Vendor support	HDInsight Cost



HDInsight Cluster Architecture



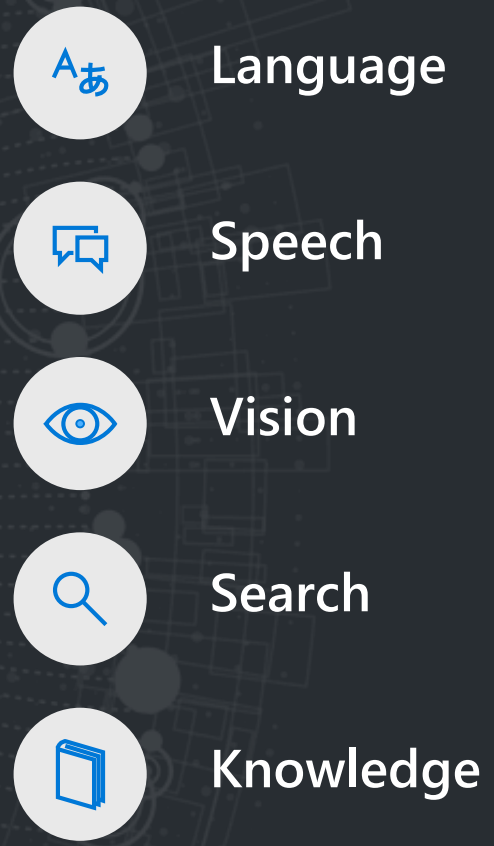
Intelligence Infrastructure

Azure Cognitive Services

Add the power of machine learning to any app

Process images, video, speech, language and more

Simple REST APIs



Microsoft AI Portfolio

People

Agent

Cortana

Applications

Office 365

Dynamics 365

SwiftKey

Pix

Customer Service and Support

Skype

Calendar.help

Services

Bot Framework

Cognitive Services

Cortana Intelligence

Cognitive Toolkit

Infrastructure

Azure Machine Learning

Azure N Series

FPGA

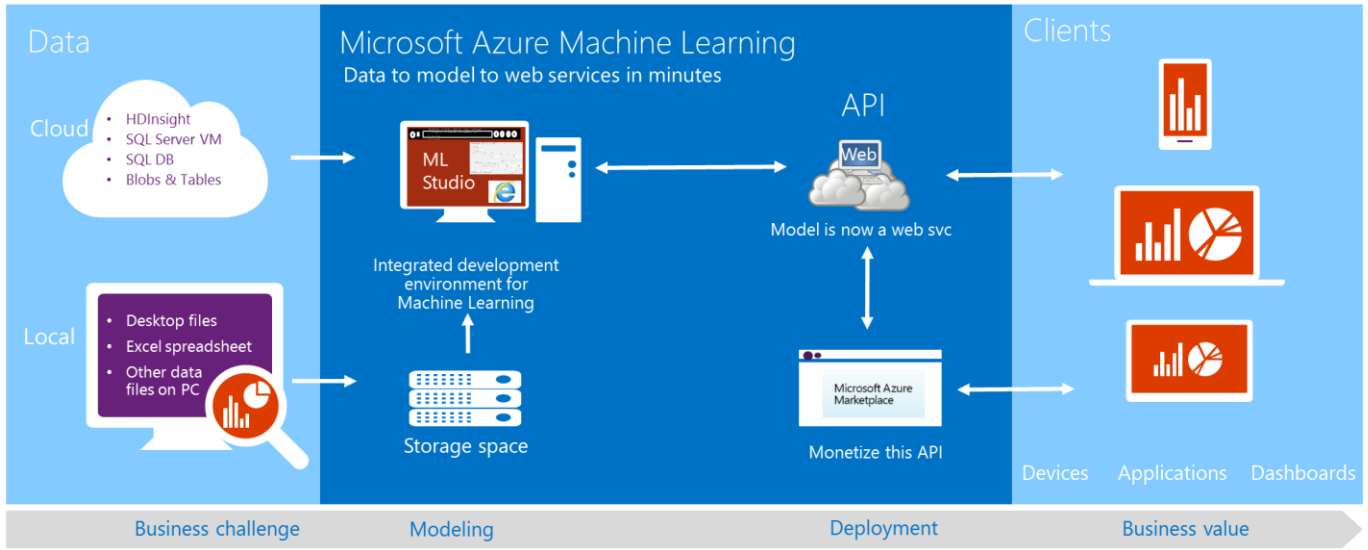
Machine Learning

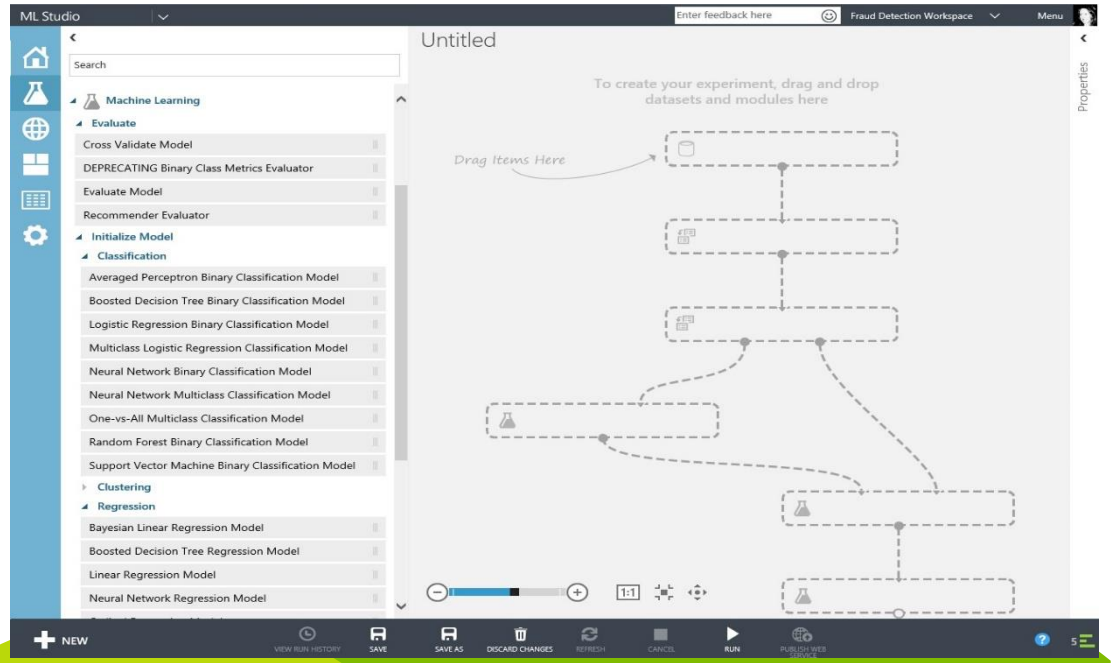
What is the feature?

Azure Machine Learning is a fully managed Platform as a Service in the cloud, integrated with data sources like HDInsight, Azure SQL Database, SQL in a VM, etc.

Based mainly on the open source language **R**, it leverages algorithms from businesses like **Bing**, **Xbox**, etc., in more than 350 packages. The APIs can be then published in the marketplace.

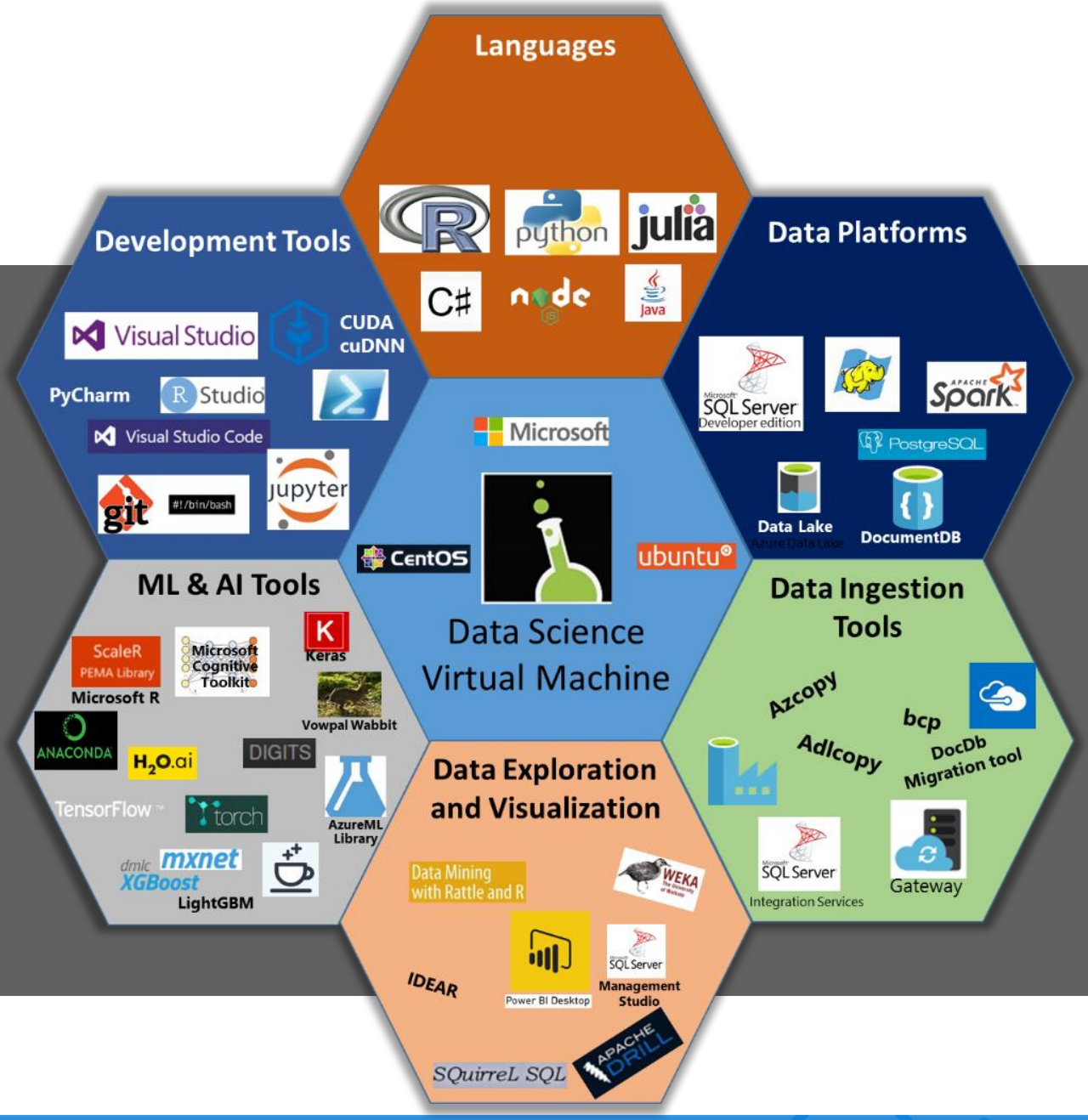
Azure ML **APIs at marketplace**: Wealth Score, Giving Score, Frequently Bought Together, Recommendations, Anomaly Detection, Lexicon Based Sentiment Analysis, Forecasting-Exponential Smoothing, etc.





Data Science VM ?

Comprehensive cloud based Data Science Environment to empower Data Scientists



Data Science VM – Current Offers

Data Science Virtual Machine

By Microsoft

Virtual machine with tools for the data science modeling and development

Software plans start at

Free

Get it now

Deep Learning toolkit for Data Science VM

By Microsoft

A deep learning tookit for the data science virtual machine

Price varies

Get it now

Data Science Virtual Machine for Linux (Ubuntu)

By Microsoft

Virtual machine with deep learning frameworks and tools for machine learning and data science

Software plans start at

Free

Test Drive

Data Science Virtual Machine for Linux (CentOS)

By Microsoft

Virtual machine with tools for the data science modeling and development

Software plans start at

Free

Get it now

Data Science Virtual Machine-Windows 2016 (Preview)

By Microsoft

Exploration, analysis, modelling, and development tools for data science

Software plans start at

Free

Get it now

VM Versions comparison – Quick Reference

Windows Edition

- ✓ Microsoft R Open with popular packages pre-installed
- ✓ Microsoft R Server Developer Edition
- ✓ Anaconda Python 2.7, 3.5 with popular packages pre-installed
- ✓ JuliaPro & Juno IDE with popular packages pre-installed
- ✓ Jupyter Notebook Server (R, Python, Julia)

- ✓ SQL Server 2016 Developer Edition: Scalable in-database analytics with R services
  - ➥ SSMS, SSIS, bcp, sqlcmd, ODBC/JDBC drivers

- ✓ IDEs and Editors
  - ➥ Visual Studio Community Edition 2015
  - ➥ Visual Studio Code
  - ➥ Azure HDInsight (Hadoop), Data Lake, SQL Server Data tools
  - ➥ Node.js, Python, and R tools for Visual Studio
  - ➥ RStudio Desktop

- ✓ Microsoft Excel 2016 *(via. Office 2016 – Windows Server 2016 only – Office ProPlus License Required)*
- ✓ Power BI desktop - (BI Dashboard Design & Analysis)

- ✓ Machine Learning Tools
  - ➥ Integration with Azure Machine Learning
  - ➥ Microsoft Cognitive toolkit (CNTK) - (deep Learning/AI)
  - ➥ Xgboost (popular ML tool in data science competitions)
  - ➥ Vowpal Wabbit (fast online learner)
  - ➥ Weka
  - ➥ Rattle (visual quick-start data and analytics tool)
  - ➥ MXnet (deep learning/AI)
  - ➥ Tensorflow
  - ➥ CUDA, CUDNN, Nvidia Drivers

- ✓ SDKs to access Azure and Cortana Intelligence Suite of services
- ✓ Azure Storage Explorer, CLI, PowerShell, AdlCopy (Azure Data Lake), AzCopy, dtui (for DocumentDB), Microsoft Data Management Gateway
- ✓ Git & GitBash, Visual Studio Team Services plugin + Windows port of most popular Linux/Unix command-line utilities accessible through GitBash/command prompt
- ✓ Apache Drill

Linux Editions

- ✓ Microsoft R Open with popular packages pre-installed
- ✓ Microsoft R Server Developer Edition (MicrosoftML not Available)
- ✓ Anaconda Python 2.7, 3.5 with popular packages pre-installed
- ✓ JuliaPro & Juno IDE with popular packages pre-installed
- ✓ JupyterHub: Multi-user Jupyter notebooks
  - ➥ R, Python, Julia, PySpark, SparkR, [also Sparkmagic on Ubuntu only]

- ✓ PostgreSQL, SQuirreL SQL (querying tool), SQL Server drivers, and command line (bcp, sqlcmd)

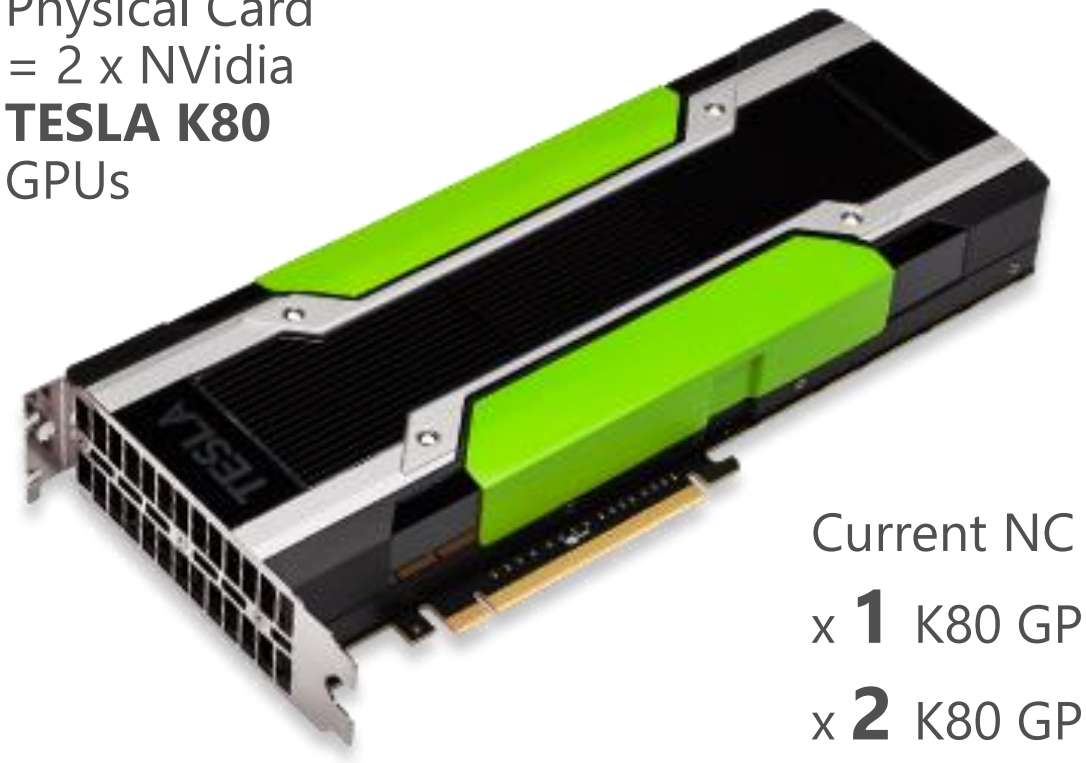
- ✓ IDEs and editors
  - ➥ Visual Studio Code
  - ➥ Vim & Emacs (with ESS, auctex) gedit
  - ➥ IntelliJ IDEA
  - ➥ RStudio Desktop & Rstudio Server
  - ➥ PyCharm
  - ➥ Atom

- ✓ Machine Learning Tools
  - ➥ Integrations with Azure Machine Learning
  - ➥ Microsoft Cognitive toolkit (CNTK)-(deep Learning/AI)
  - ➥ Xgboost (popular ML tool in data science competitions)
  - ➥ Vowpal Wabbit (fast online learner)
  - ➥ Weka
  - ➥ Rattle (visual quick-start data and analytics tool)
  - ➥ MXnet (deep learning/AI)
  - ➥ Tensorflow
  - ➥ CUDA, CUDNN, Nvidia Drivers
  - ➥ Caffe & Caffe 2
  - ➥ Torch
  - ➥ Theano
  - ➥ Keras
  - ➥ Nvidia Digits

- ✓ SDKs to access Azure and Cortana Intelligence Suite of services
- ✓ Tools for data movement and management of Azure and Big Data resources: Azure Storage Explorer, CLI
- ✓ Git
- ✓ Apache Drill
- ✓ Apache Spark - local instance (Standalone)

The Deep Learning toolkit for DSVM + Ubuntu & Windows Server 2016 Click Here

Physical Card
= 2 x NVidia
TESLA K80
GPUs



The Toolkit deploys on **NC** class Azure VMs with GPUs

Great for **Deep Learning** workloads

Current NC Class VM SKU Configurations:

x **1** K80 GPU	- 1/2 Physical Card	- **12** GB GDDR5 VRAM
x **2** K80 GPUs	- 1 Physical Card	- **24** GB GDDR5 VRAM
x **4** K80 GPUs	- 2 Physical Cards	- **48** GB GDDR5 VRAM

DEMO

Resources

Azure HDInsight (Hadoop on Cloud)
https://docs.microsoft.com/en-in/azure/hdinsight/

Microsoft Cognitive Services
https://docs.microsoft.com/en-us/azure/cognitive-services/welcome
https://azure.microsoft.com/en-in/services/cognitive-services/face/

Microsoft Cognitive Services (Model Customization)
https://customvision.ai/
https://azure.microsoft.com/en-us/services/cognitive-services/custom-speech-service/

Microsoft Pricing Calculator
https://azure.microsoft.com/en-in/pricing/calculator/

Microsoft Container Services
https://azure.microsoft.com/en-in/overview/containers/

Microsoft Serverless Computing
https://docs.microsoft.com/en-in/azure/azure-functions/

Microsoft Azure Market Place
https://azuremarketplace.microsoft.com/en-us/marketplace/

Provision the Windows Data Science Virtual Machine on Azure
https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/provision-vm

Microsoft Azure Machine Learning Studio
https://docs.microsoft.com/en-in/azure/machine-learning/studio/

Microsoft Azure Machine Learning Cheat Sheet
https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-cheat-sheet

Thank You!

Post your queries/doubts on ***https://elearn.bits-pilani.ac.in/*** website