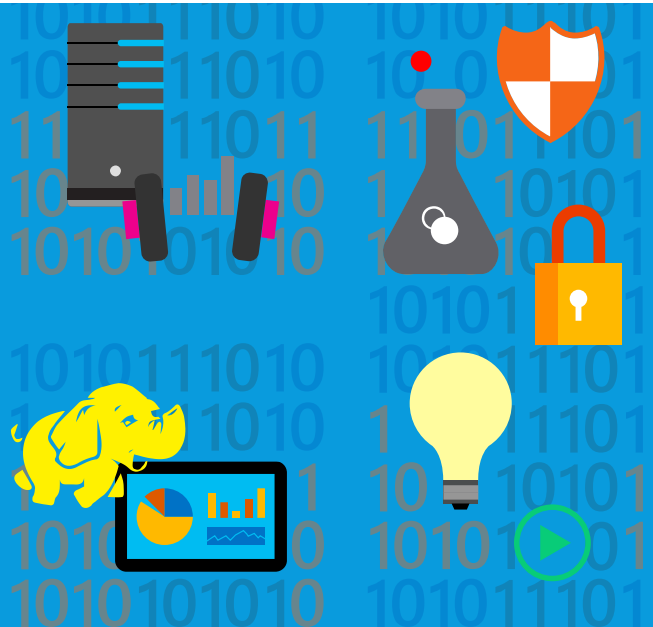
DATA SCIENCE AND BIG DATA OVERVIEW

Azure Machine Learning Track

Vladimír Mužný & Marek Chmel

2017 Global Azure BOOTCAMP



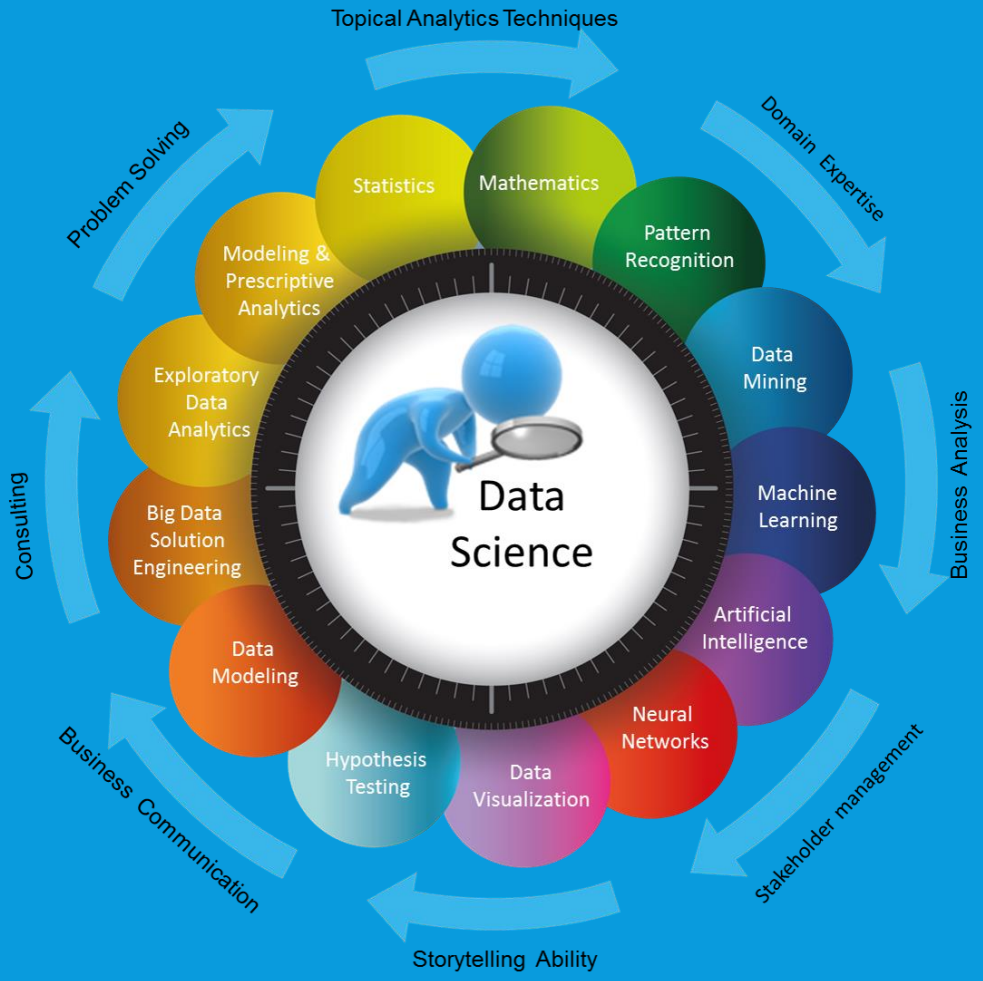
ÚVODEM

- Úvod do Data Science
- Big Data a Data Science
- Práce s daty
  - Získávání dat
  - Analýza
  - Vizualizace

WHAT IS DATA SCIENCE

- Is the study of the generalizable extraction of knowledge from data (Wikipedia)
- Is getting predictive and/or actionable insight from data (Neil Raden)
- Involves extracting, creating, and processing data to turn it into business value. – Vincent Granville (Developing Analytic Talent: Becoming a Data Scientist)

- **Just because you know how the tools work, doesn't mean you are doing data science!**

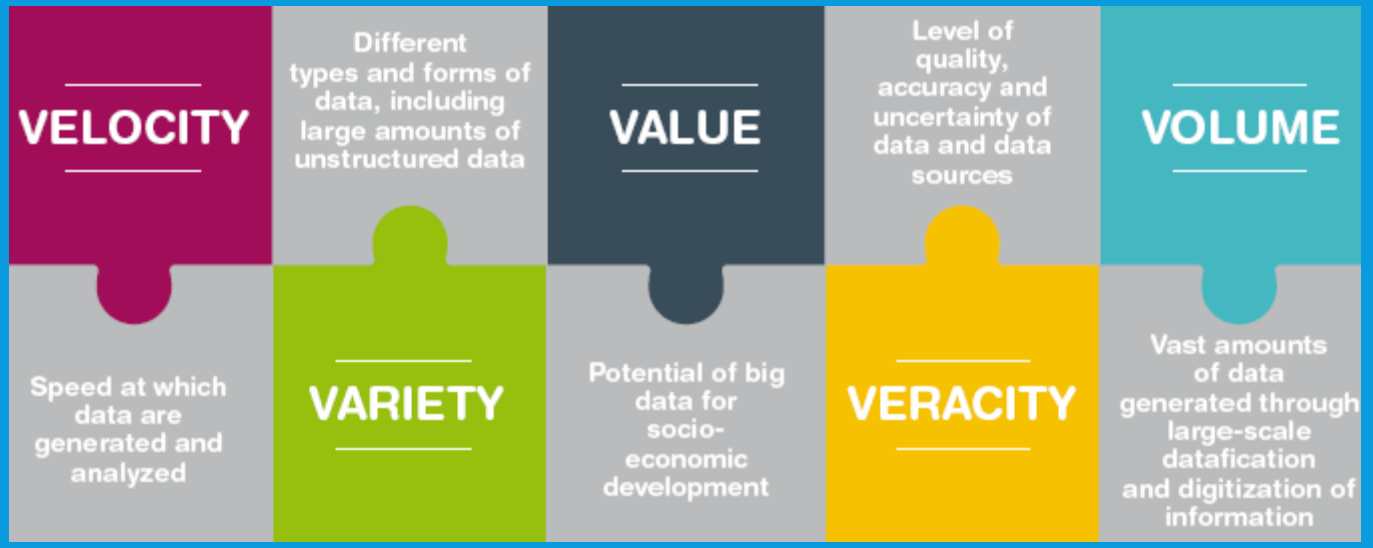


WHAT IS BIG DATA

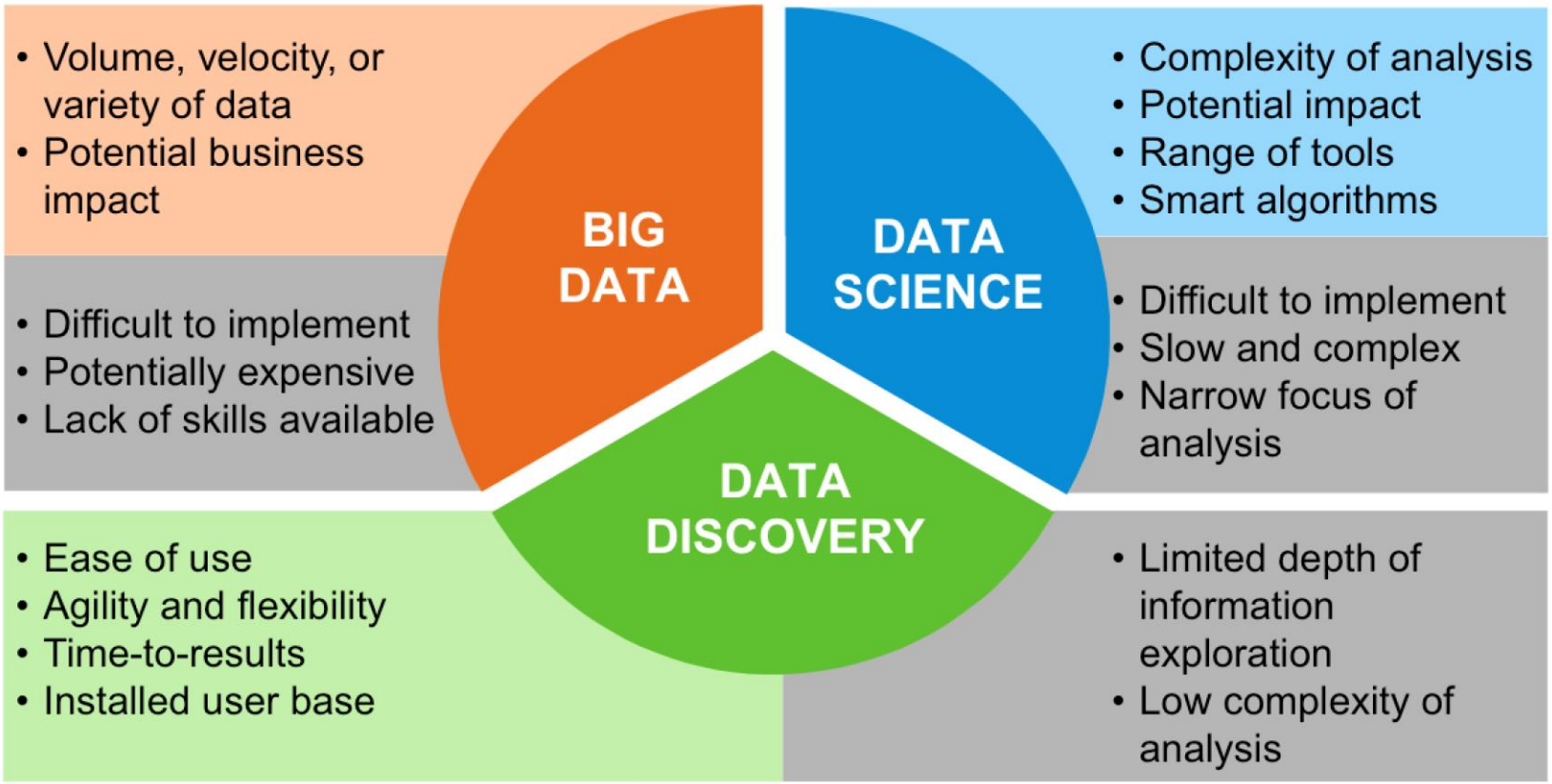
Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Challenges include capture, storage, analysis, data curation, search, sharing, transfer, visualization, querying, updating and information privacy.

5Vs OF BIG DATA

- Big Data is a big thing. It will change our world completely and is not a passing fad that will go away. To understand the phenomenon that is big data, it is often described using five Vs: Volume, Velocity, Variety, Veracity and Value



BIG DATA VS DATA SCIENCE



- Data science is a field that comprises of everything that is related to data cleansing, preparation and analysis

- Big data is something that can be used to analyze insights which can lead to better decisions and strategic business moves

- Data analytics involves automating insights into a certain dataset as well as purposes the usage of queries and data aggregation procedures

DATA SCIENCE ROLES

Data Scientist

A highly educated and skilled person who can solve complex data problems by employing deep expertise in scientific disciplines (mathematics, statistics or computer science)

Data Professional

A skilled person who creates or maintains data systems, data solutions or implements predictive modelling.

Roles: Database Administrator, Database Developer, or BI Developer

Software Developer

A skilled person who designs and develops programming logic, and can apply machine learning to integrate predictive functionality into applications

DATASCIENCE CERTIFICATION BY MICROSOFT

- https://academy.microsoft.com
- 15 different courses to choose from, you need to pass 10
- Practical exam based on AzureML experiment with Cortana Analytics Suite
- MCP Exam 70-774 "Perform Cloud Data Science with Azure Machine"
- MCP Exam 70-773 "Analyzing Big Data with Microsoft R"
- MCP Exam 70-774 "Perform Cloud Data Science with Azure Machine Learning"
- MCP Exam 40-475 "Designing and Implementing Big Data Analytics Solutions"

REAL WORLD APPLICATION FOR DS/BD

Sensor data analysis

Buyer propensity models

Social network analysis

Predictive maintenance

Search engine optimization

Churn analysis

Natural resource exploration

Weather forecasting

Healthcare outcomes

Fraud detection

Life sciences research

Targeted advertising

Network intrusion detection

Smart meter monitoring

DATA SCIENCE DEMYSTIFIED

- Data science is used in many fields today but is still clouded with many myths
- We would like to uncover come of those to explain what big data and data science is about

M #1 - BIG DATA WILL INSTANTLY YIELD GREAT INSIGHTS

- Needs
- Evaluation

- Architecture
- Procurement

- Installation
- Security

- Development
- Use

M #2 - BIG DATA WILL MAKE ANALYTICS EASIER

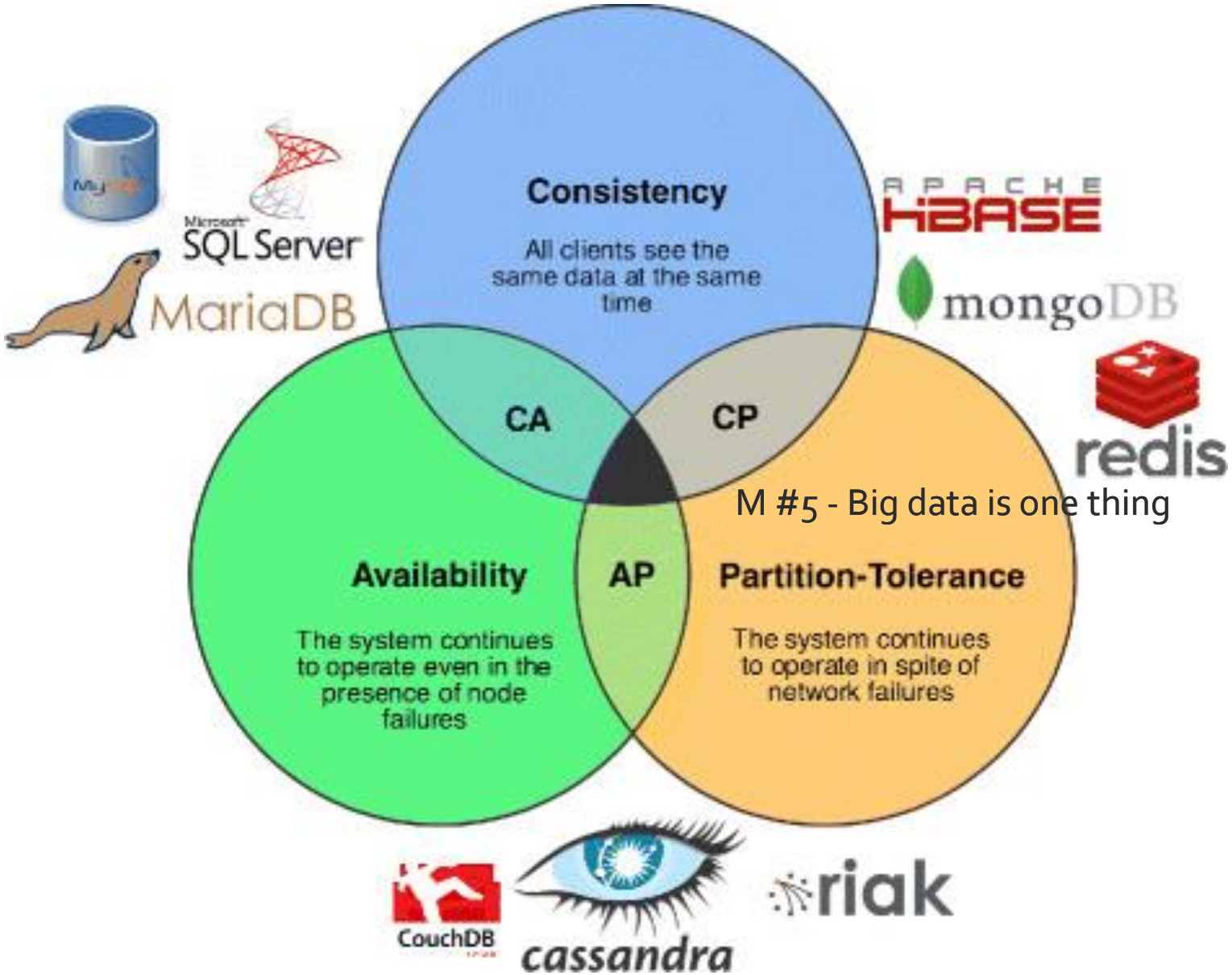
- Big data will bring more data sources to the game and that will not make analytics easier at all. Users will just have more sources for data preparation, more tools for data processing and in the end even more tools for reporting.

M #3 - BIG DATA WILL BE EASY TO SET UP

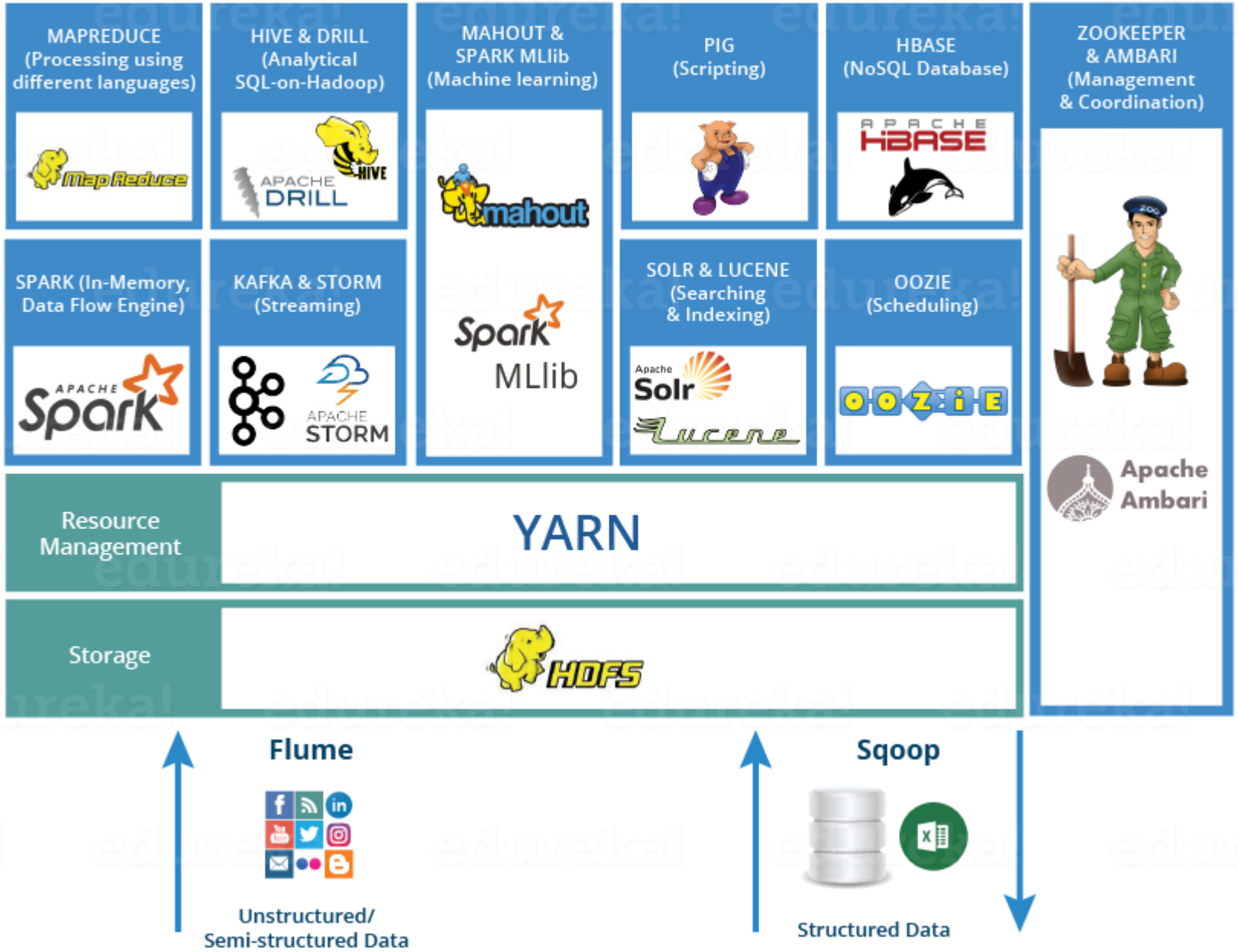
- Cloud setup considerations
  - Services or machines
  - Response times / backup
  - Zoning
  - Access methods
- On-Premises considerations
  - HW
  - Physical locations
  - Physical security
  - Network Access

M #4 - BIG DATA WILL REPLACE MY RELATIONAL DATABASES

- it is impossible for a distributed computer system to simultaneously provide more than two out of three of the following guarantees



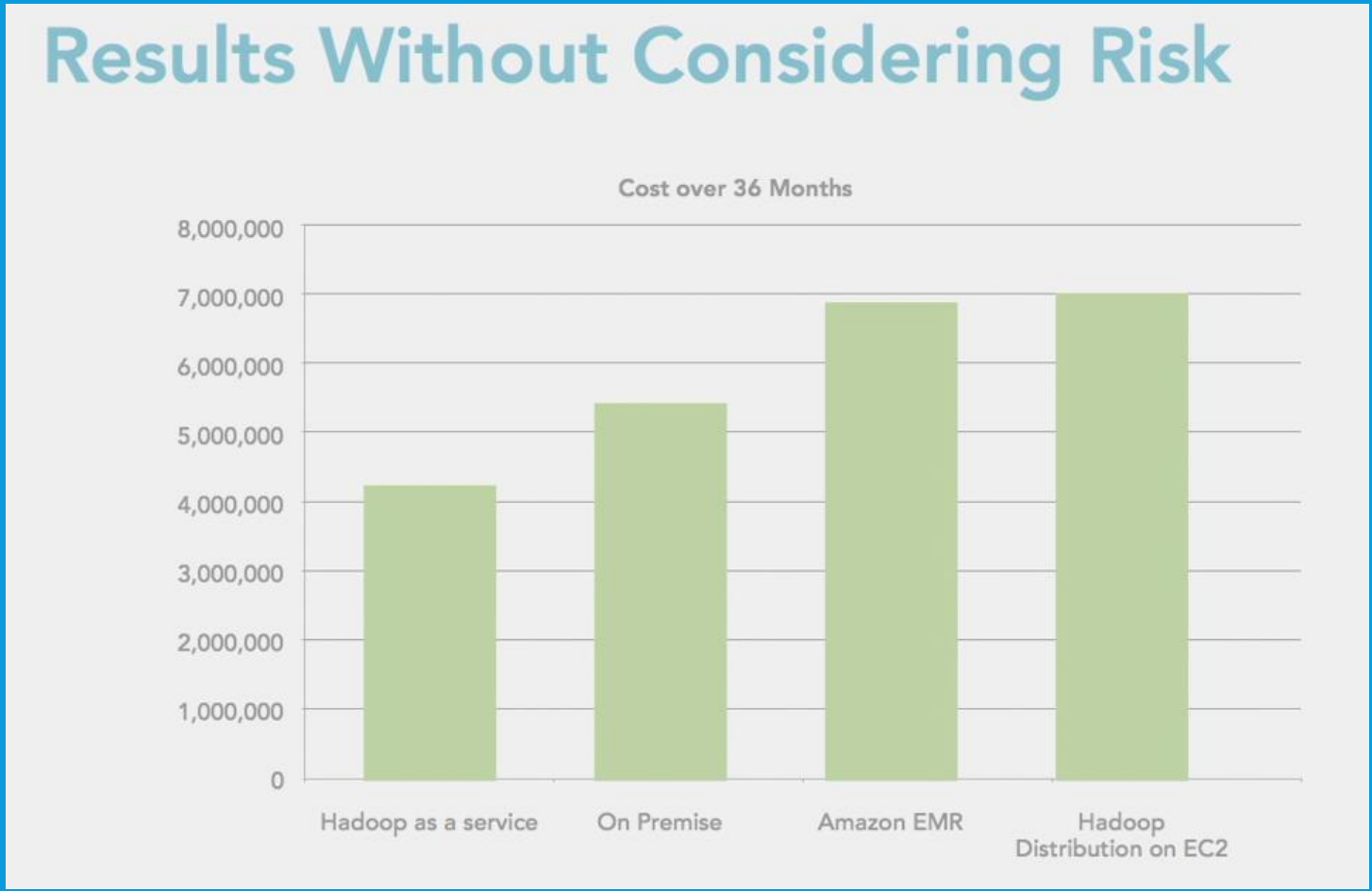
M #5 - Big data is one thing



M #5 - BIG DATA IS ONE THING

Hadoop is a framework, consisting of many different tools which can be used for data processing, storage, automation etc.

M #6 - BIG DATA IS CHEAP



M #7 - BIG DATA SYSTEMS ARE FAST

- Apache Hadoop framework used for distributed storage and processing of big data sets using the MapReduce programming model.
- Speed challenges
  - Dependency on MapReduce
  - Distributed data
  - Low-power disk
  - Conversion of analytic functions

M #8 - BIG DATA IS THE IT TEAM'S JOB

- Cloudová technologie pro
  - Příjem a přípravu dat
  - Transformaci dat
  - Publikování a použití dat v následné analýze
- Dostupná jako „Data Factory“ („Datové továrny“ v české verzi Azure Portal)
- Publikovaná v US regionech a v North Europe regionu

CHASM

M #9 - BIG DATA IS NECESSARY

PROS

- Real-time data
- Unstructured data
- All the data
- New product offerings

CONS

- Lots of setup
- Unfamiliar interface
- Volumes can be hard to comb through

DATA ACQUISITION

- Variable quality of data, different type of data
  - Cleansing, transforming
- Getting data via traditional ETL
- Data Inspection
  - Profiling task
  - PowerPivot

CATEGORIES OF DATA

- Numeric
  - Discrete
  - Continuous
- Categorical
- Ordinal

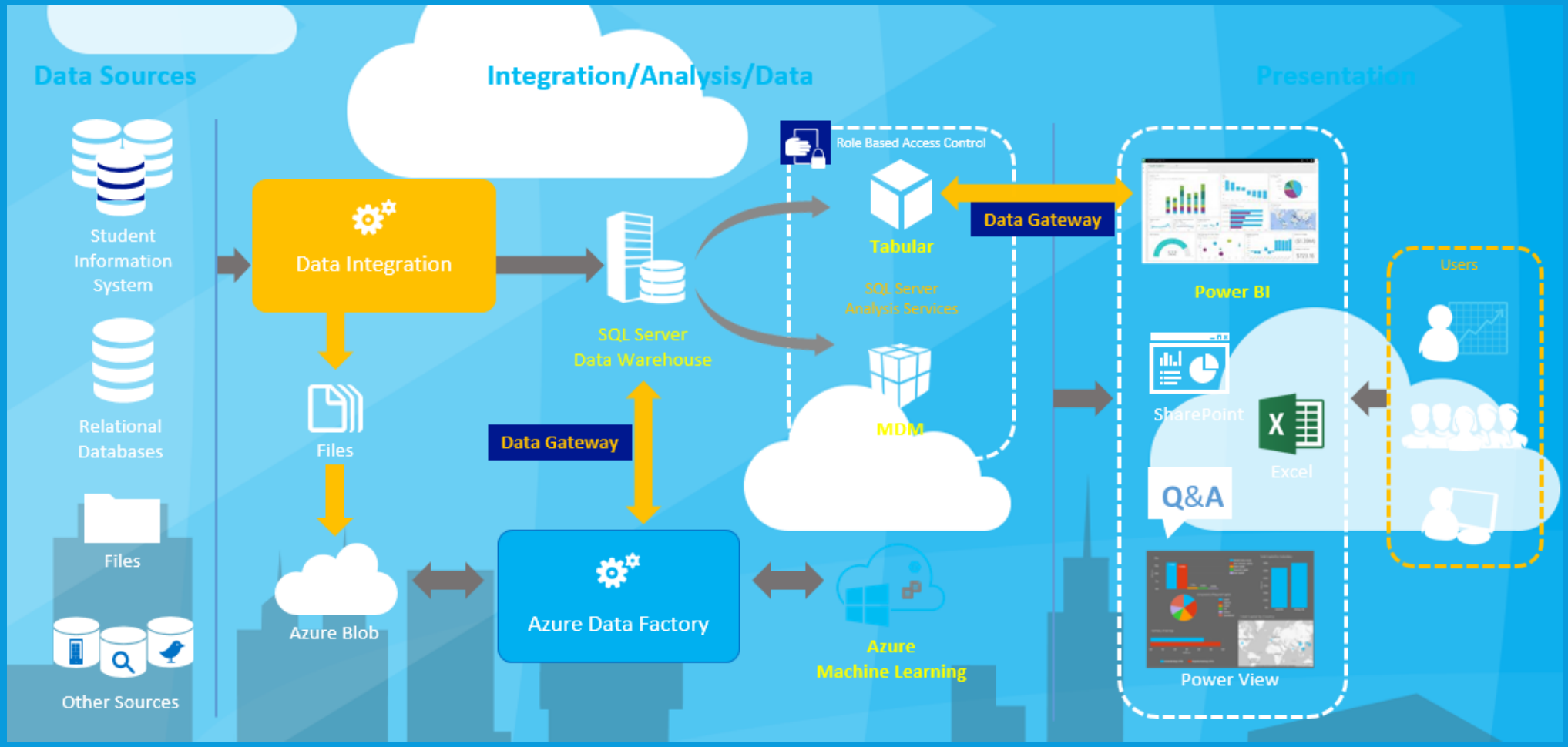
DESCRIPTIVE STATISTICS

- Key words
  - Mean, Median, Variance
- Visualization is the key

DATA ANALYSIS

- Quantiles
- Variance
- Correlations
- Histograms
- ANOVA
- Z-Test

BIGGER PICTURE



ADDITIONAL READING

- Free eBook
  - https://blogs.msdn.microsoft.com/microsoft_press/2015/04/15/free-ebook-microsoft-azure-essentials-azure-machine-learning/
- Free Azure trial offer at:
  - http://azure.microsoft.com/en-us/pricing/free-trial
- Free Azure Machine Learning Trial offer at:
  - https://studio.azureml.net/Home
- Azure Machine Learning:
  - http://azure.microsoft.com/en-us/services/machine-learning/
- Azure Machine Learning Data Market:
  - http://datamarket.azure.com/browse?query=machine%20learning
- Azure Machine Learning gallery
  - https://gallery.azureml.net
- Azure Machine Learning blog
  - http://blogs.technet.com/b/machinelearning
- Videos: PASS Data Science Virtual Chapter
  - https://www.youtube.com/channel/UCqB3xWdwjA9soFV6EOu7qfg