Microsoft
Technology
Centers
Experience the
Microsoft Cloud

Data Platform Overview
Azure ML Data Camp

Ivan Kosyakov
MTC Architect, Ph.D.

Microsoft

Business Trends

Top Manager

IT

Analyst

Business User

Data Scientist

Big Data (3v)	Strategic Analytics	Business Unit Analytics

End User Analytics	Advanced Analytics
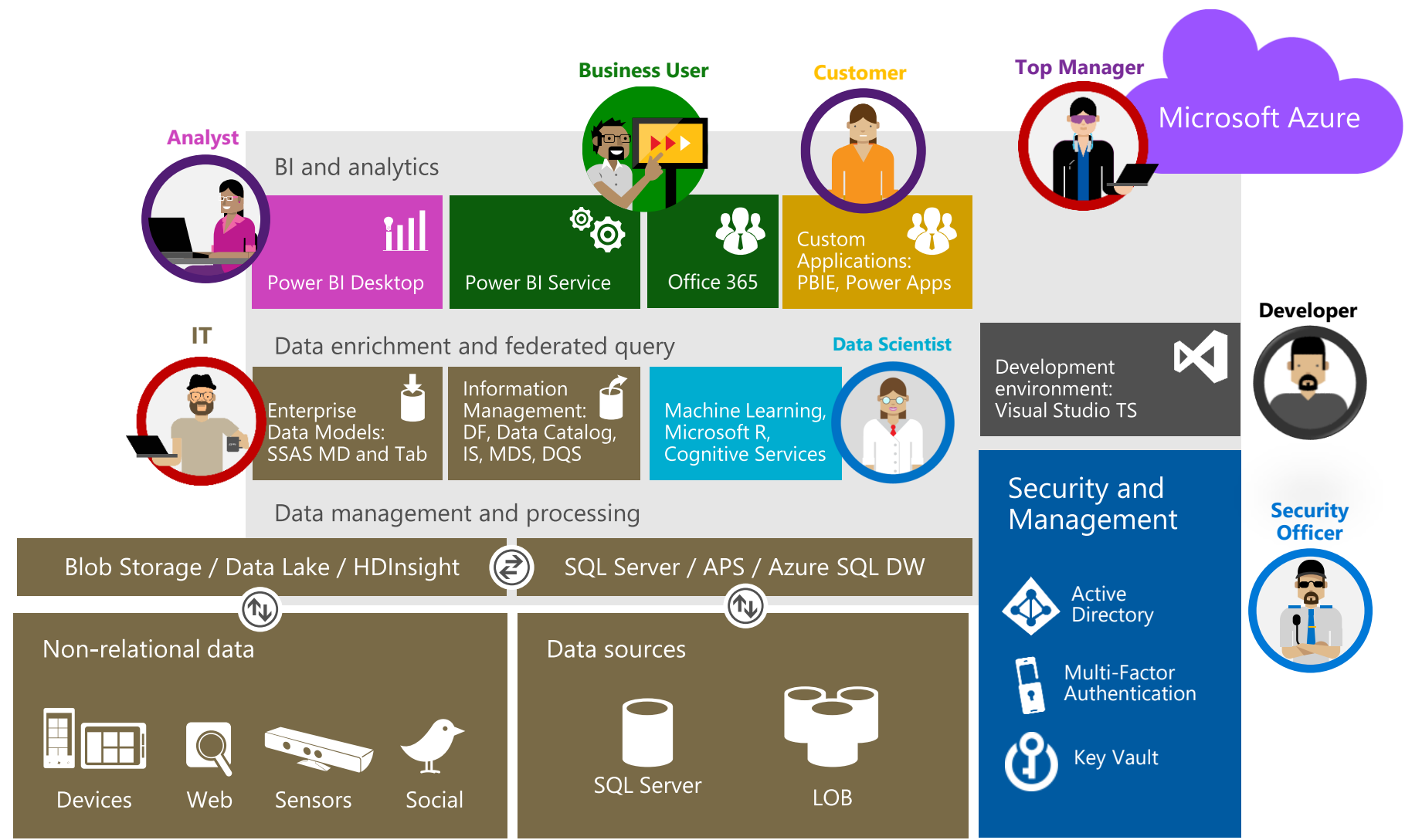
Cortana Analytics Suite

Unparalleled security

>$1B annual spend on security & ecosystem

Hyper scalability

100+ datacenters in over 40 countries

Most comprehensive

Over 16 services to support advanced analytics

- Preconfigured Solutions
- Dashboards and Visualizations
- Machine Learning and Analytics
- Big Data Store
- Information Management

+

- Personal Digital Assistant – Cortana
- Perceptual Intelligence

Intelligent by design

$12B R&D spend annually
1000+ scientists and engineers
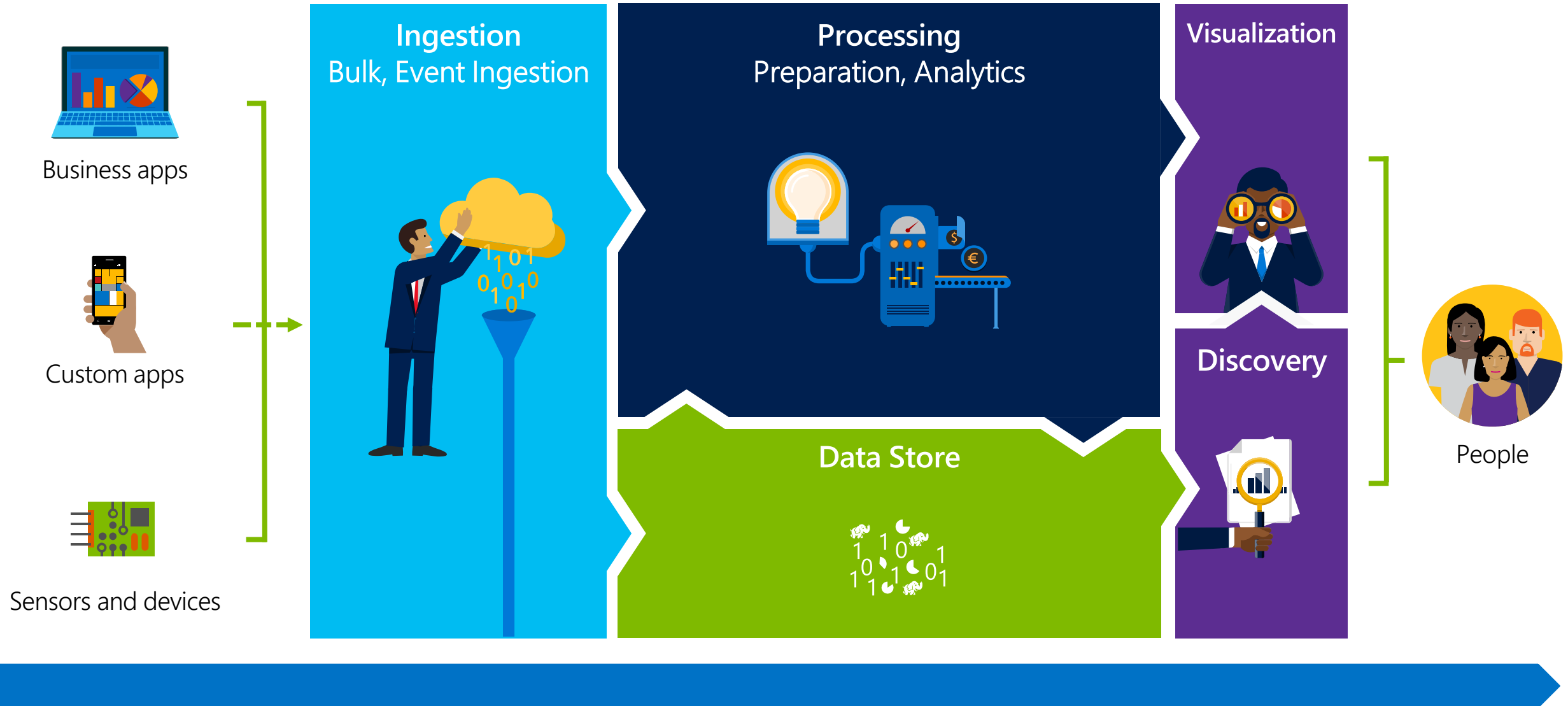**10,000+ peer reviewed publications**

Face APIs
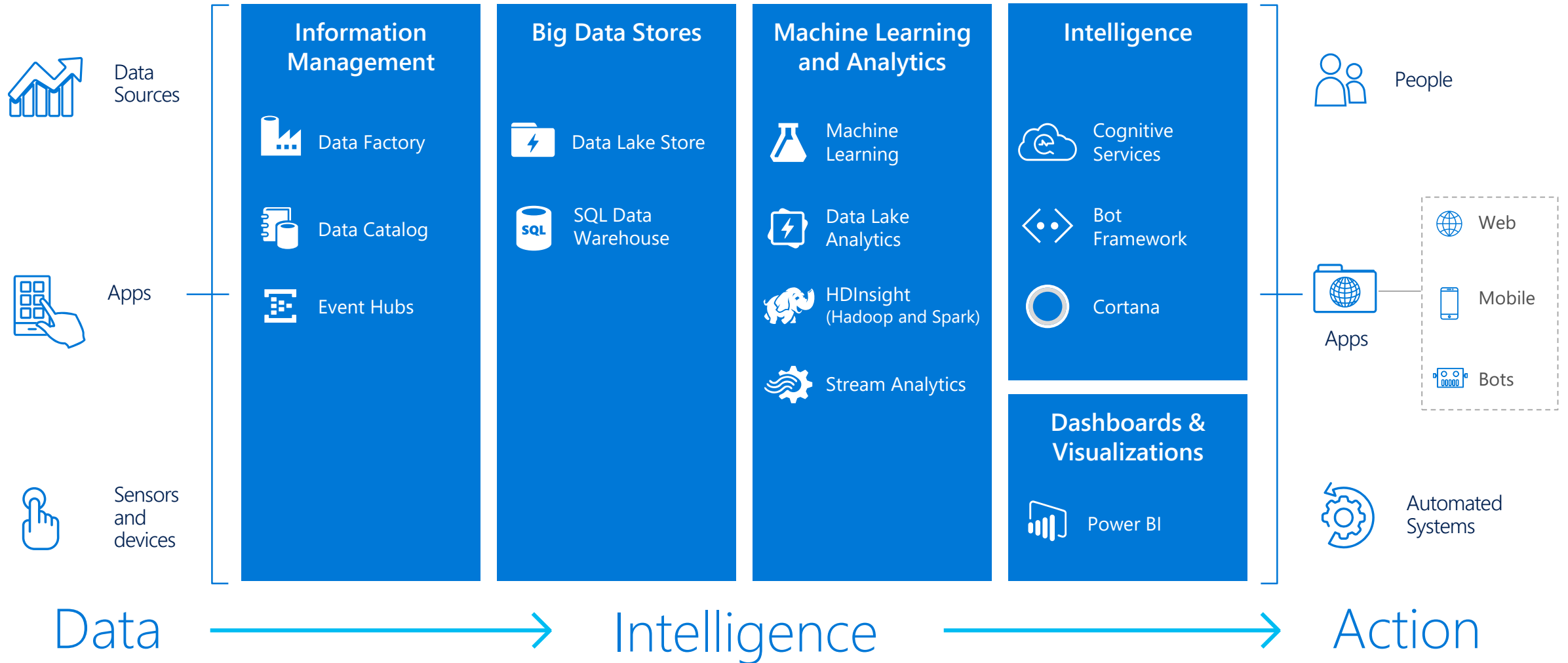
Speech APIs

Computer Vision APIs

Text Analytics

Transform data into intelligent action in the cloud

Big Data Flow
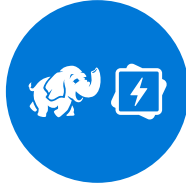
Example reference architecture
Analyze Data Stored in Azure

- Track realtime data from IOT Suite: collect data from IOT Suite in permanent store (ADLS)

- Track other Azure data (Azure Website generating web logs) and store in ADLS

- Run Machine Learning through R Server for HDInsight to find patterns in data

- Show results in BI tools (Power BI)

Reference architecture for Big Data

| Data Generation | Real-Time | Storage | Processing | Visualizations |

Data Generation: Azure Website, Devices, Crawlers, Bots, Sensors, Facebook, Twitter

Real-Time: Azure Event Hubs, Azure Stream Analytics, Spark Streaming for Azure HDInsight, Storm for Azure HDInsight

Storage: Azure SQL DB, Azure SQL DW, Azure Document DB, Azure Data Lake Store, HBase on Azure HDInsight

Processing: SQL Server Analysis Services, Azure Machine Learning, Azure Data Lake Analytics, Azure HDInsight, Hive, Spark SQL, R Server, Sqoop, Pig, Spark MLib, Oozie

Visualizations: Azure Marketplace, SSRS, SharePoint BI, Excel BI, Power BI, Datazen

Azure Data Factory

Big Data Decision Tree

Big Data is often described as a solution to the "three V's problem", and how we choose right solution depends on which one of these problems we are trying to solve first:

- **Volume**: need to store and query hundreds of terabytes of data or more, and the total volume is growing. Processing systems must be scalable to handle increasing volumes of data, typically by scaling out across multiple machines.
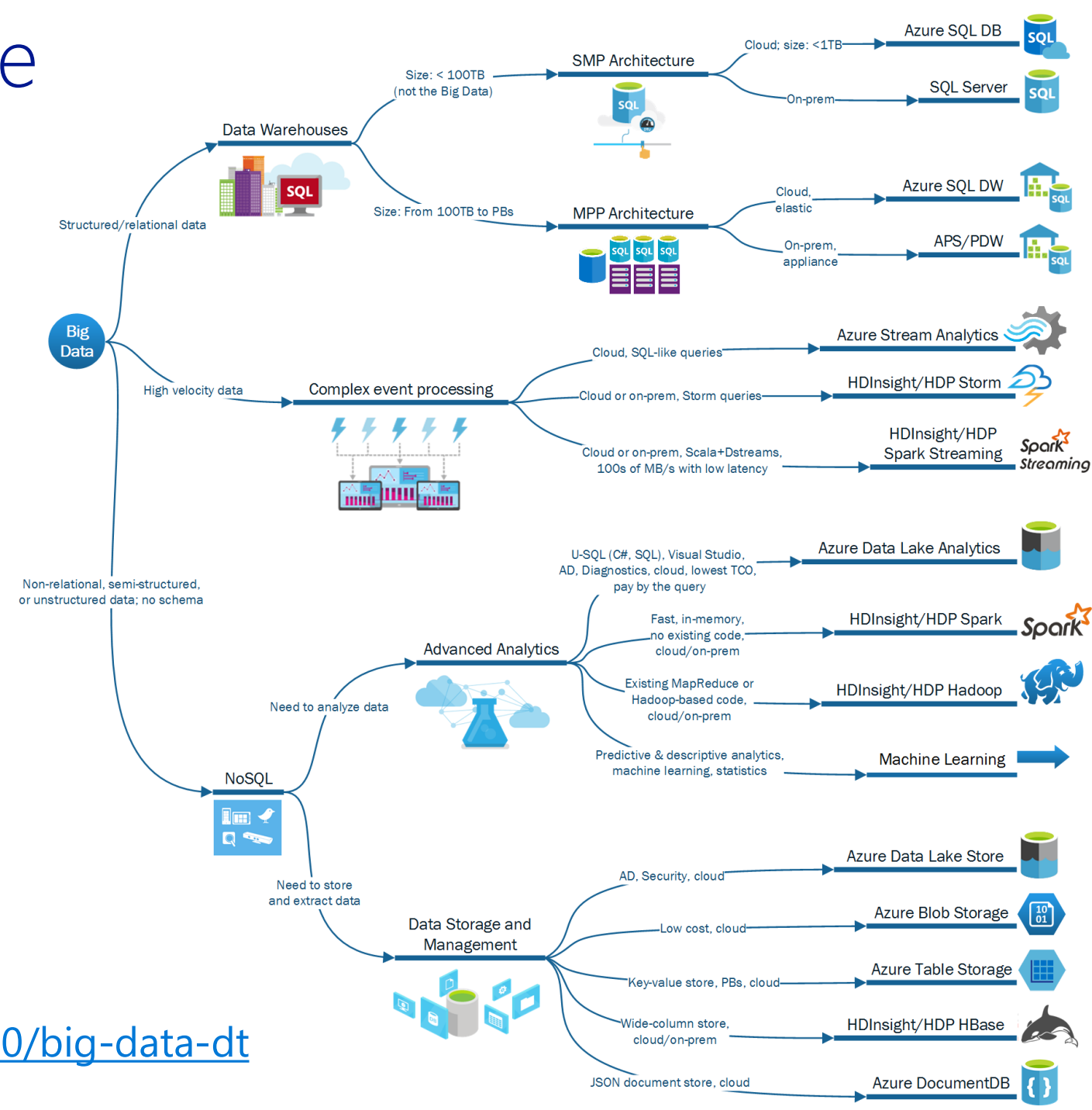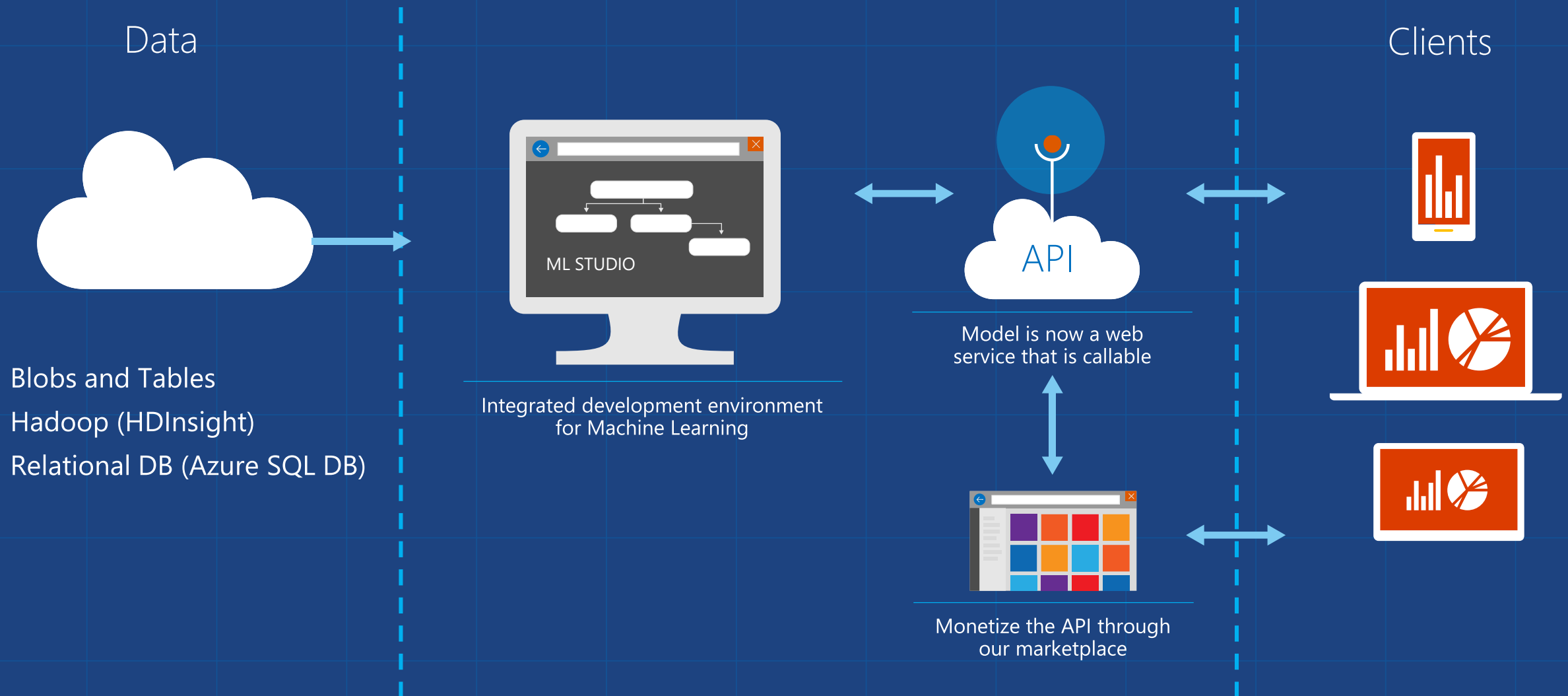
- **Velocity**: need to collect data at an increasing rate from many new types of devices, from a fast-growing number of users, and from an increasing number of devices and applications per user. Processing systems must be able to return results within an acceptable timeframe, often almost in real-time.

- **Variety**: situation when data do not match any existing data schema – semi-structured or unstructured data.

Source: https://biz-excellence.com/2016/08/30/big-data-dt

Decision Tree

Big Data

Structured/relational data → Data Warehouses
- Size: < 100TB (not the Big Data) → **SMP Architecture**
  - Cloud; size: <1TB → Azure SQL DB
  - On-prem → SQL Server
- Size: From 100TB to PBs → **MPP Architecture**
  - Cloud, elastic → Azure SQL DW
  - On-prem, appliance → APS/PDW

High velocity data → Complex event processing
- Cloud, SQL-like queries → Azure Stream Analytics
- Cloud or on-prem, Storm queries → HDInsight/HDP Storm
- Cloud or on-prem, Scala+Dstreams, 100s of MB/s with low latency → HDInsight/HDP Spark Streaming

Non-relational, semi-structured, or unstructured data; no schema → NoSQL

- Need to analyze data → **Advanced Analytics**
  - U-SQL (C#, SQL), Visual Studio, AD, Diagnostics, cloud, lowest TCO, pay by the query → Azure Data Lake Analytics
  - Fast, in-memory, no existing code, cloud/on-prem → HDInsight/HDP Spark
  - Existing MapReduce or Hadoop-based code, cloud/on-prem → HDInsight/HDP Hadoop
  - Predictive & descriptive analytics, machine learning, statistics → Machine Learning

- Need to store and extract data → **Data Storage and Management**
  - AD, Security, cloud → Azure Data Lake Store
  - Low cost, cloud → Azure Blob Storage
  - Key-value store, PBs, cloud → Azure Table Storage
  - Wide-column store, cloud/on-prem → HDInsight/HDP HBase
  - JSON document store, cloud → Azure DocumentDB

Azure Machine Learning Service

Data -> Predictive model -> Operational web API in minutes

Data

Clients

ML STUDIO

Integrated development environment
for Machine Learning

API

Model is now a web
service that is callable

Monetize the API through
our marketplace

Blobs and Tables

Hadoop (HDInsight)
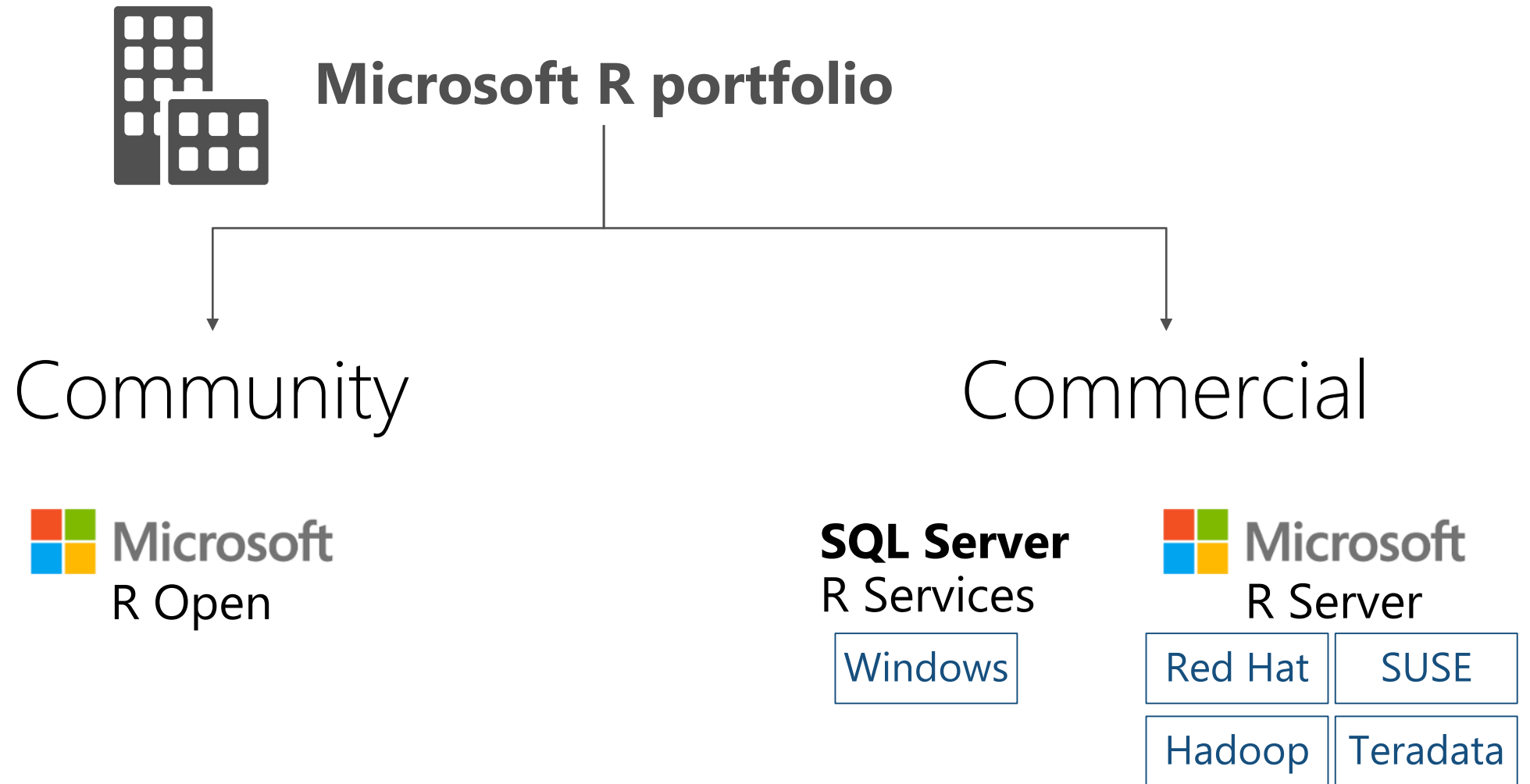
Relational DB (Azure SQL DB)

Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.

ANOMALY DETECTION

One-class SVM — >100 features, aggressive boundary

PCA-based anomaly detection — Fast training

CLUSTERING

K-means

MULTI-CLASS CLASSIFICATION

Fast training, linear model → **Multiclass logistic regression**

Accuracy, long training times → **Multiclass neural network**

Accuracy, fast training → **Multiclass decision forest**

Accuracy, small memory footprint → **Multiclass decision jungle**

Depends on the two-class classifier, see notes below → **One-v-all multiclass**

REGRESSION

Ordinal regression — Data in rank ordered categories

Poisson regression — Predicting event counts

Fast forest quantile regression — Predicting a distribution

Linear regression — Fast training, linear model

Bayesian linear regression — Linear model, small data sets

Neural network regression — Accuracy, long training time

Decision forest regression — Accuracy, fast training

Boosted decision tree regression — Accuracy, fast training, large memory footprint

START

- Discovering structure → Clustering
- Finding unusual data points → Anomaly detection
- Three or more / Predicting categories → Multi-class classification
- Two / Predicting categories → Two-class classification
- Predicting values → Regression

TWO-CLASS CLASSIFICATION

Two-class SVM — >100 features, linear model

Two-class averaged perceptron — Fast training, linear model

Two-class logistic regression — Fast training, linear model

Two-class Bayes point machine — Fast training, linear model

Accuracy, fast training → **Two-class decision forest**

Accuracy, fast training, large memory footprint → **Two-class boosted decision tree**

Accuracy, small memory footprint → **Two-class decision jungle**

>100 features → **Two-class locally deep SVM**

Accuracy, long training times → **Two-class neural network**

Microsoft R portfolio

Machine Learning Decision Tree

Source: https://biz-excellence.com/2016/09/13/machine-learning-dt