

od STUDENTA *do* SPECJALISTY BUSINESS *konferencja* INTELLIGENCE

SQL Server 2016
w świecie Big Data

Bartłomiej Graczyk
Data Platform Solution Architect
bartlomiej.graczyk@microsoft.com
bartek@graczyk.info.pl

Bartłomiej GRACZYK

Architekt rozwiązań SQL | DW | BI | MS Licensing Expert



- Wieloletnie doświadczenie w realizacji projektów w środowiskach komercyjnych, jak i administracji publicznej z zakresu rozwiązań Microsoft, w szczególności Business Intelligence & Data Warehouse / EPM / BPM
- Ekspert w zakresie licencjonowania produktów Microsoft oraz Software Asset Management
- Autor publikacji i prelegent na konferencjach i sesjach tematycznych (m.in. Time For SharePoint, VirtualStudy.pl, Microsoft SQL Server 2008 R2 Launch, SQL Server 2014 Wpływa na ocean Big Data, szkolenia MVA)
- Microsoft Certified Trainer
- Most Valuable Professional (MVP) w kategorii SQL Server (2013-2015)
- Od października 2015 - Data Platform Solution Architect, Microsoft

Data Platform Solution Architect
bagra@microsoft.com
bartek@graczyk.info.pl





Agenda

Dane na świecie –
"wczoraj,
dzisiaj, jutro..."

Wyzwania oraz
zmieniające się
potrzeby w zakresie
przetwarzania danych

SQL Server
2016 w świecie
Big Data



Agenda

Dane na świecie –
"wczoraj,
dzisiaj, jutro..."



Dane na świecie

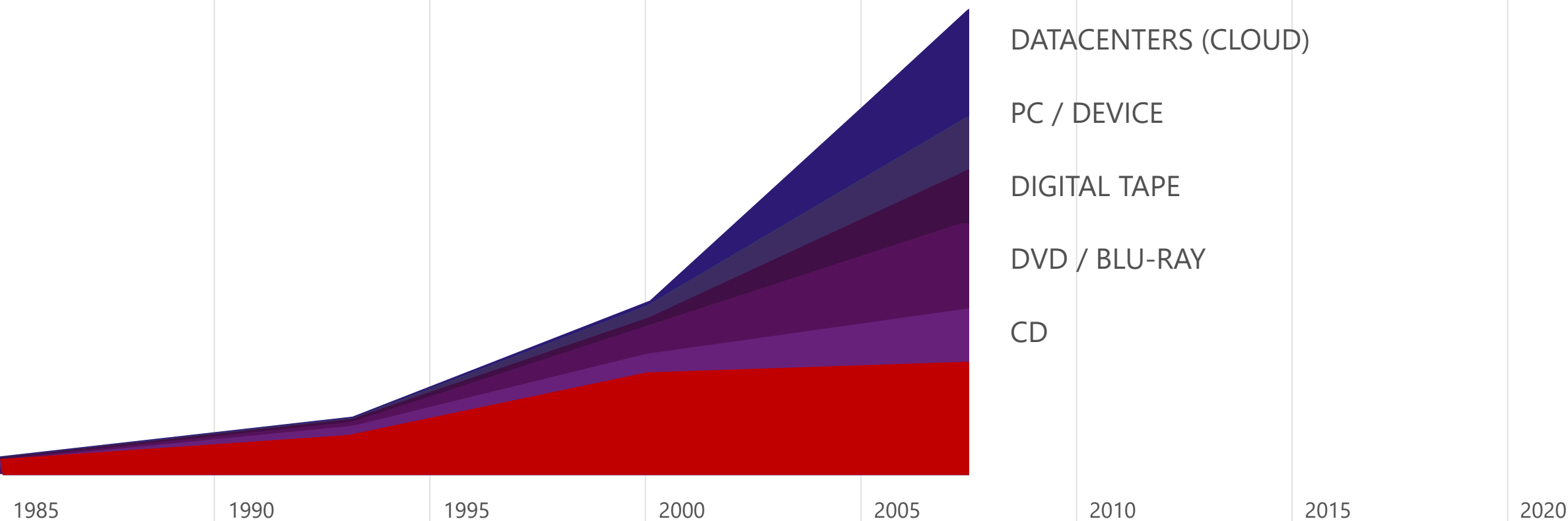
 CYFROWE
 ANALOGOWE





Dane na świecie

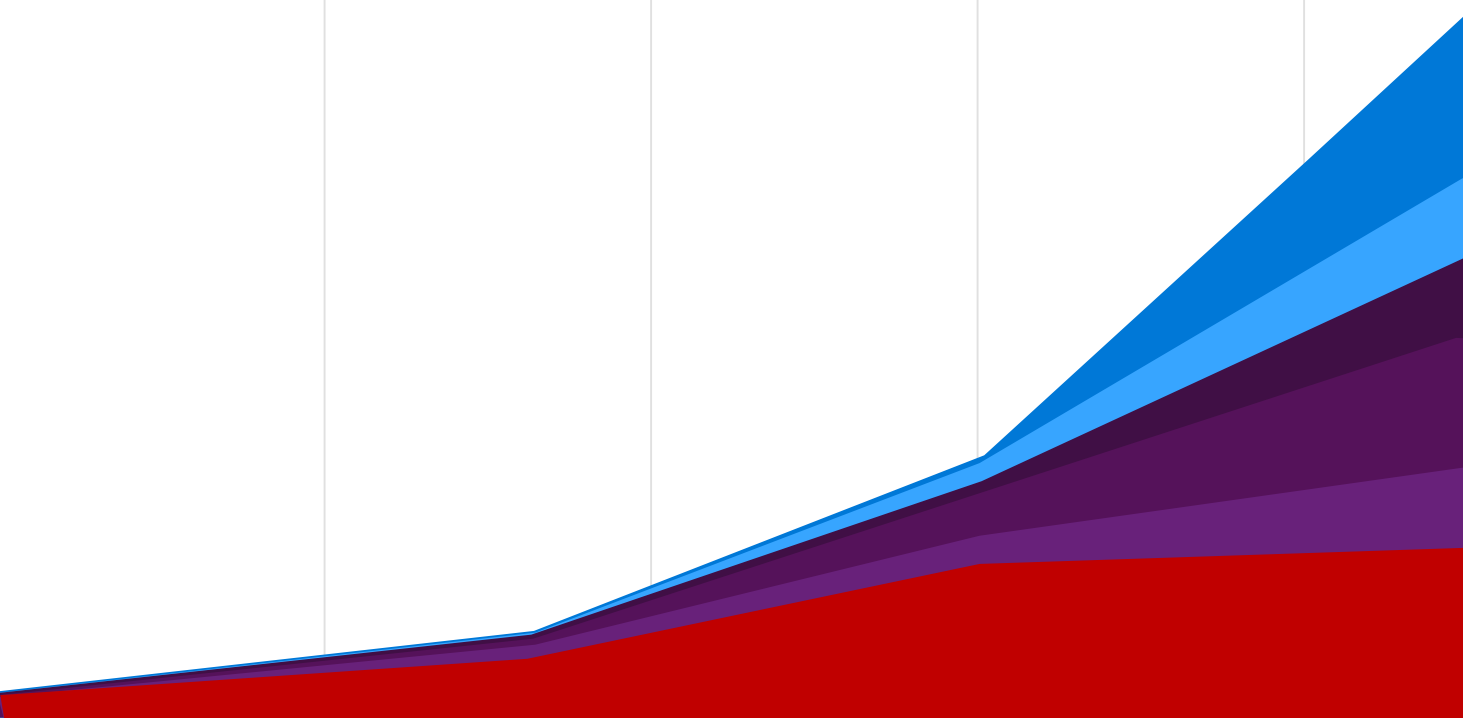
 CYFROWE
 ANALOGOWE





Dane na świecie

- POŁĄCZONE
- CYFROWE
- ANALOGOWE



DATACENTERS (CLOUD)

PC / DEVICE

DIGITAL TAPE

DVD / BLU-RAY

CD

1985

1990

1995

2000

2005

2010

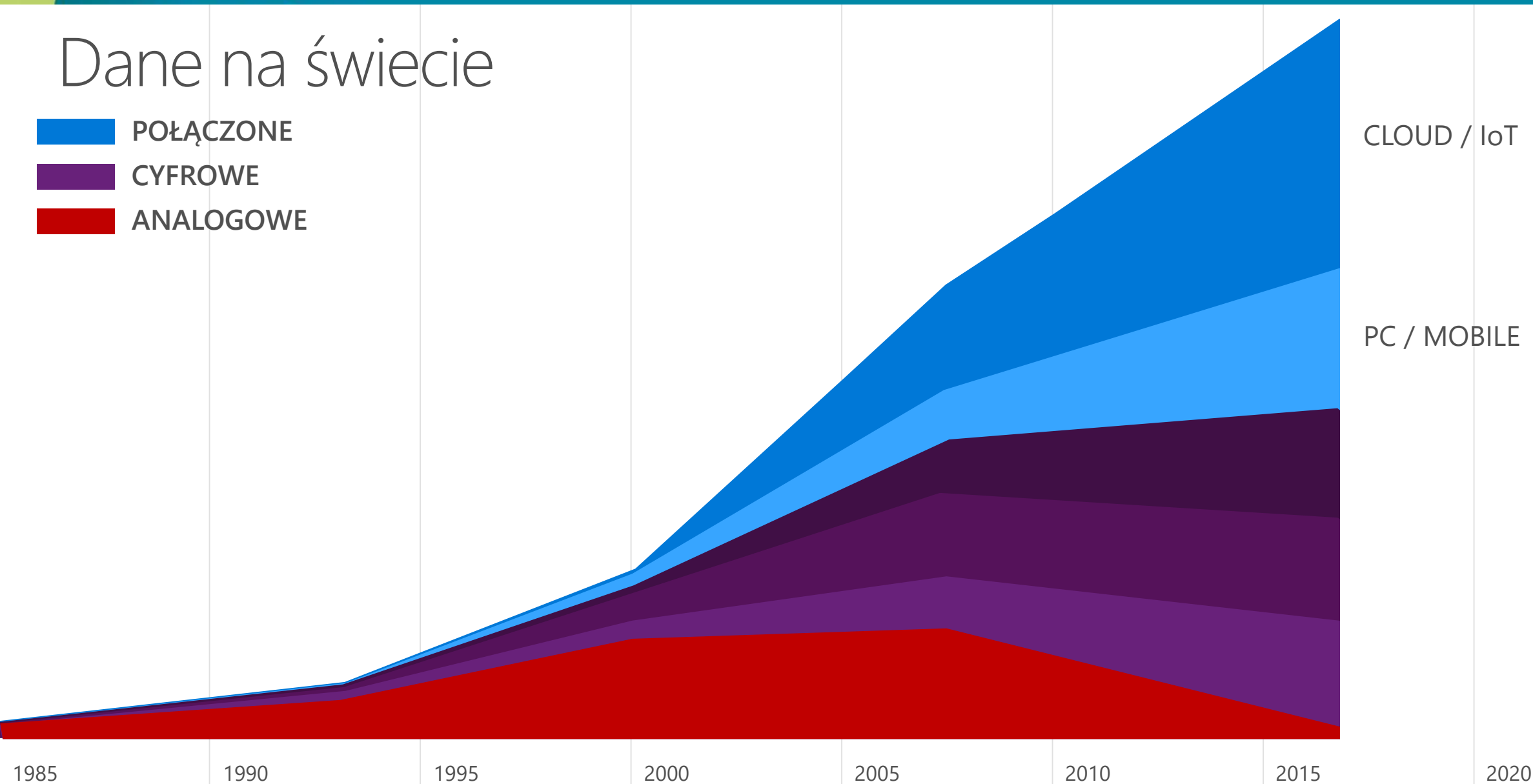
2015

2020



Dane na świecie

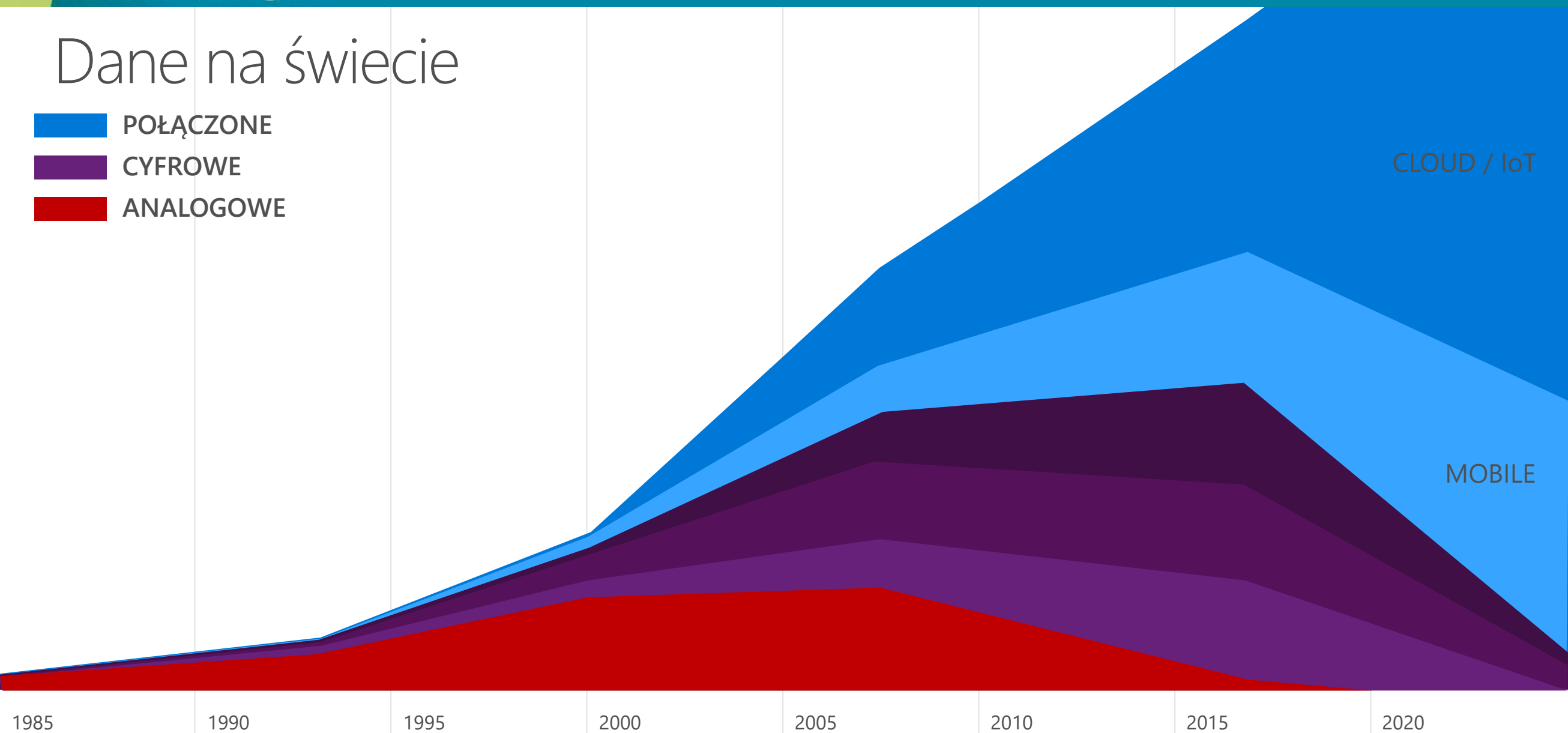
- POŁĄCZONE
- CYFROWE
- ANALOGOWE





Dane na świecie

- POŁĄCZONE
- CYFROWE
- ANALOGOWE





40 ZB = 57x

**ilość wszystkich ziarenek piasku
na wszystkich plażach świata**

40 ZB = 40 trylionów gigabajtów



Agenda

Dane na świecie –
"wczoraj,
dzisiaj, jutro..."

Wyzwania oraz
zmieniające się
potrzeby w zakresie
przetwarzania danych



Zainteresowanie światem Big Data

Rosnąca ilość źródeł danych i znaczące ilości danych wymagające przetwarzania i analizy danych w różnych formatach



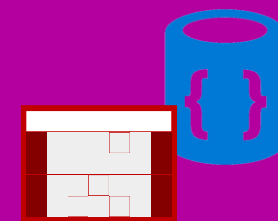
Posiadane dane są zbyt wartościowe, aby je usunąć, jednak ich przechowywanie jest jednocześnie dość kosztowne



Znaczący spadek kosztu przestrzeni dyskowej.



Wymiana danych pomiędzy systemami, w tym również bazami nierelacyjnymi



Adaptacja technologii Big Data takich jak Hadoop oraz formatów stosowanych w bazach no SQL



Agenda

Dane na świecie –
“wczoraj,
dzisiaj, jutro...”

Wyzwania oraz
zmieniające się
potrzeby w zakresie
przetwarzania danych

SQL Server
2016 w świecie
Big Data



SQL Server 2016 w świecie Big Data





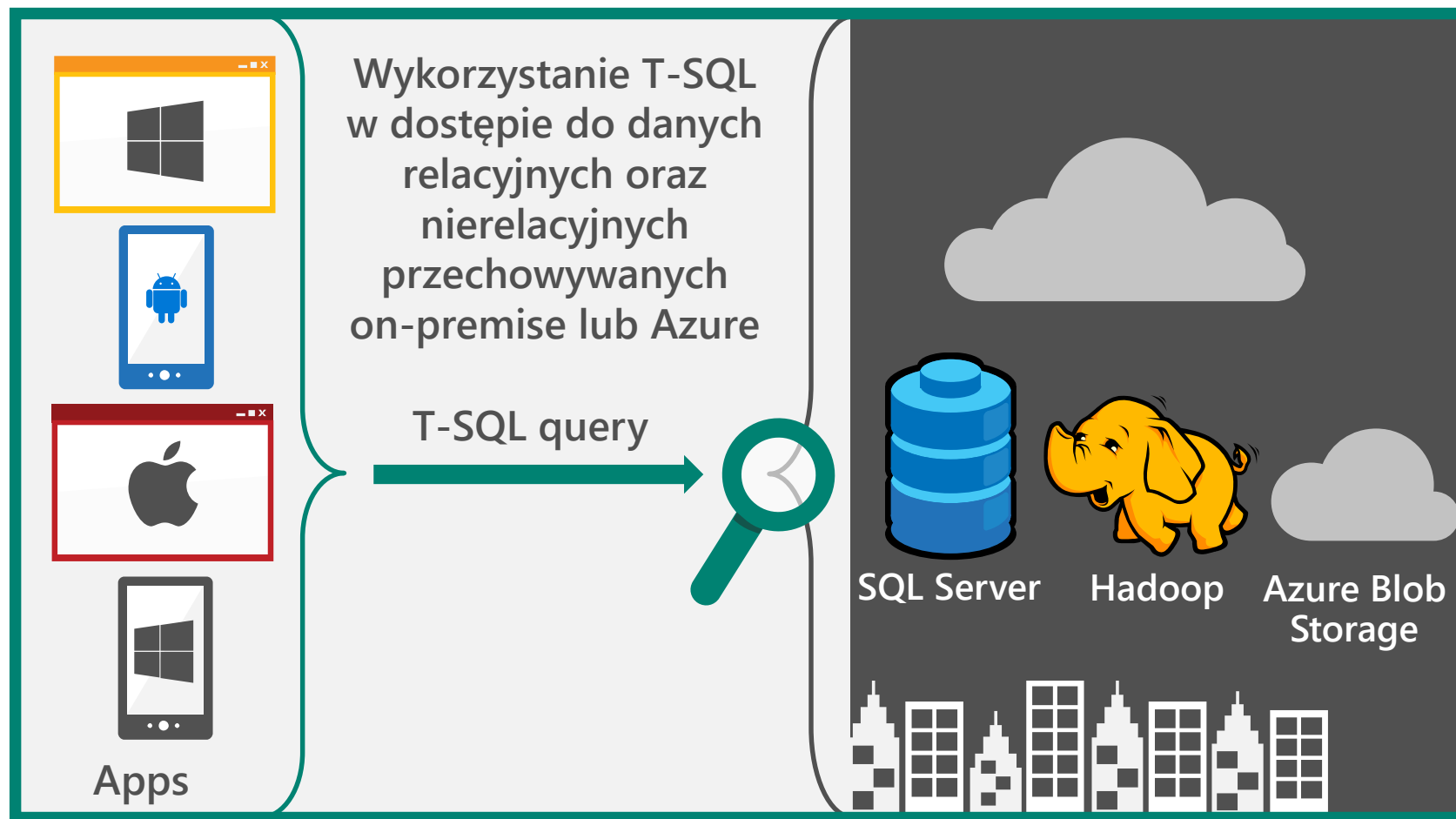
Co to jest Polybase ?



Technologia (framework) pozwalająca na integrację danych świata relacyjnego (SQL Server) oraz nierelacyjnego (Hadoop) z wykorzystaniem języka T-SQL



PolyBase w SQL Server 2016



Korzyści

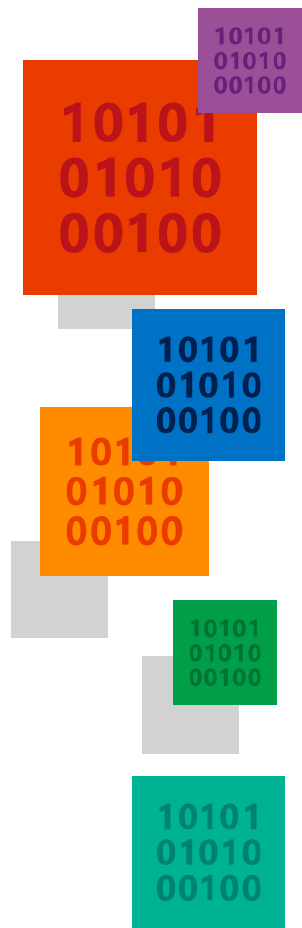
Możliwości analizy danych w oparciu o różnorodne dane

Możliwość wykorzystania dotychczasowej wiedzy oraz narzędzi w zakresie SQL Server oraz BI

Skrócony proces dostępu do danych oraz uproszczony proces ETL



Wymagania do instalacji PolyBase

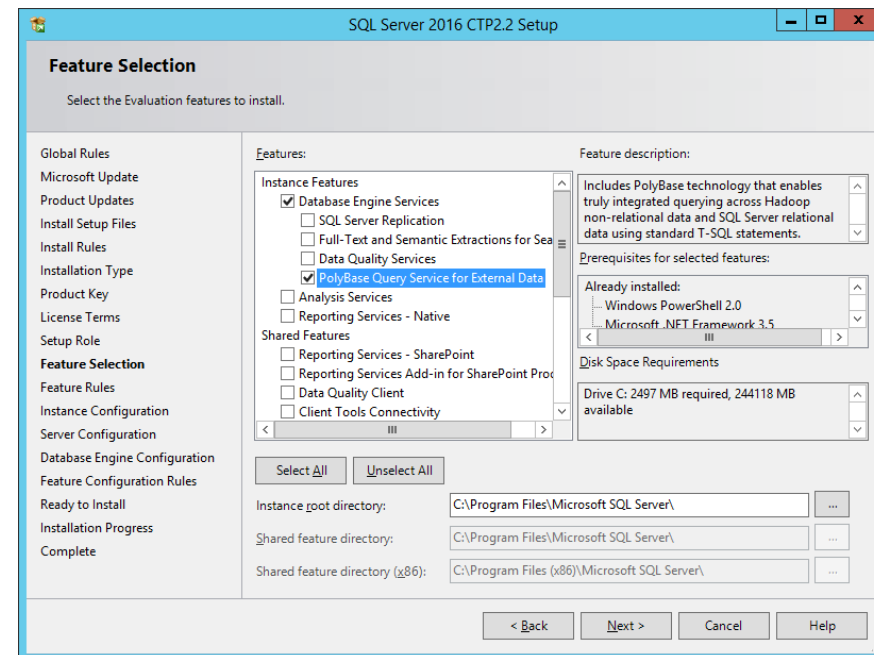


- 64-bit SQL Server Enterprise / Developer (Standard only compute node)
- Microsoft .NET Framework 4.0
- Oracle Java SE RunTime Environment (JRE) version 7.51 lub wyższy
- Pamięć minimalna: 4GB
- Minimalna przestrzeń dyskowa: 2GB



Kreator instalacji PolyBase

- Instalacja z poziomu SQL Server Installation Center (wymagany nośnik)
- W oknie usług należy wybrać PolyBase Query Service for External Data.
- Należy wskazać konto, na którym uruchomione zostaną usługi:
PolyBase Engine Service
oraz
PolyBase Data Movement Service



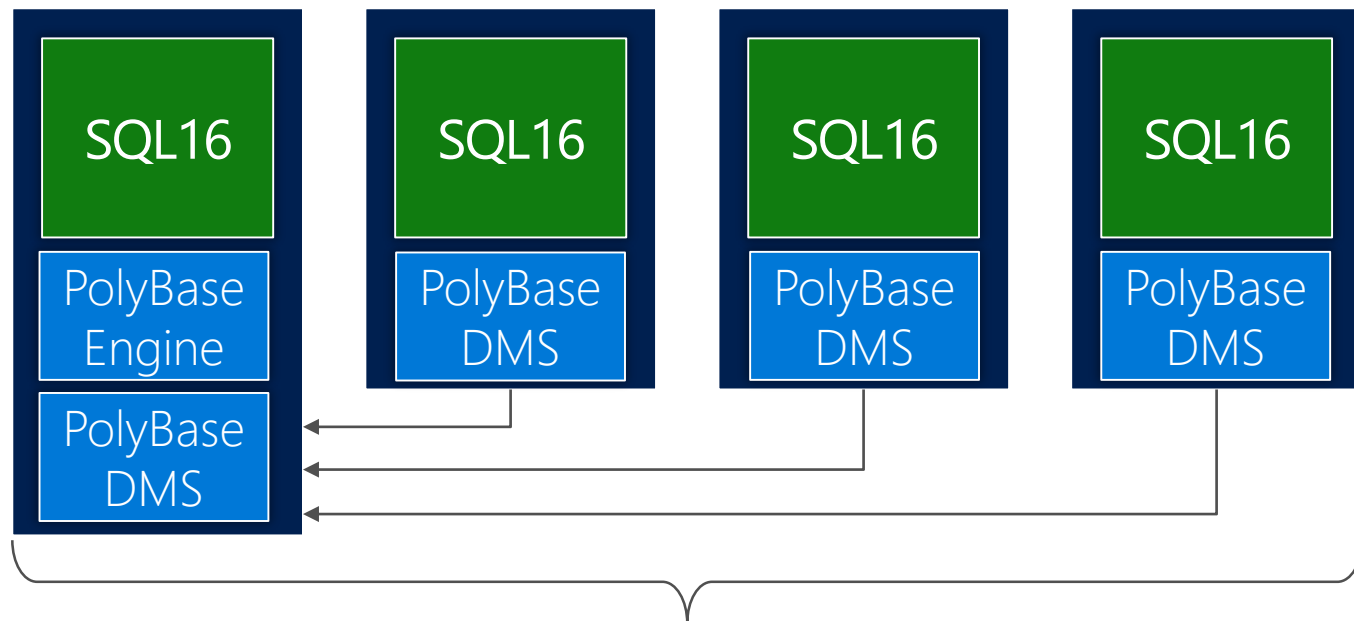


Komponenty wprowadzone w SQL Server 2016

- "PolyBase" Engine Service
- "PolyBase" Data Movement Service (HDFS Bridge)
- External Table Constructs
- MR Push-down Computation Support



Polybase instance vs PolyBase Group



-- Enter head node details:
(*head node machine name,
head node dms control channel port,
head node sql server name*)

```
EXEC sp_polybase_join_group  
'SQL-CMP01', 16450, 'MSSQLSERVER';
```

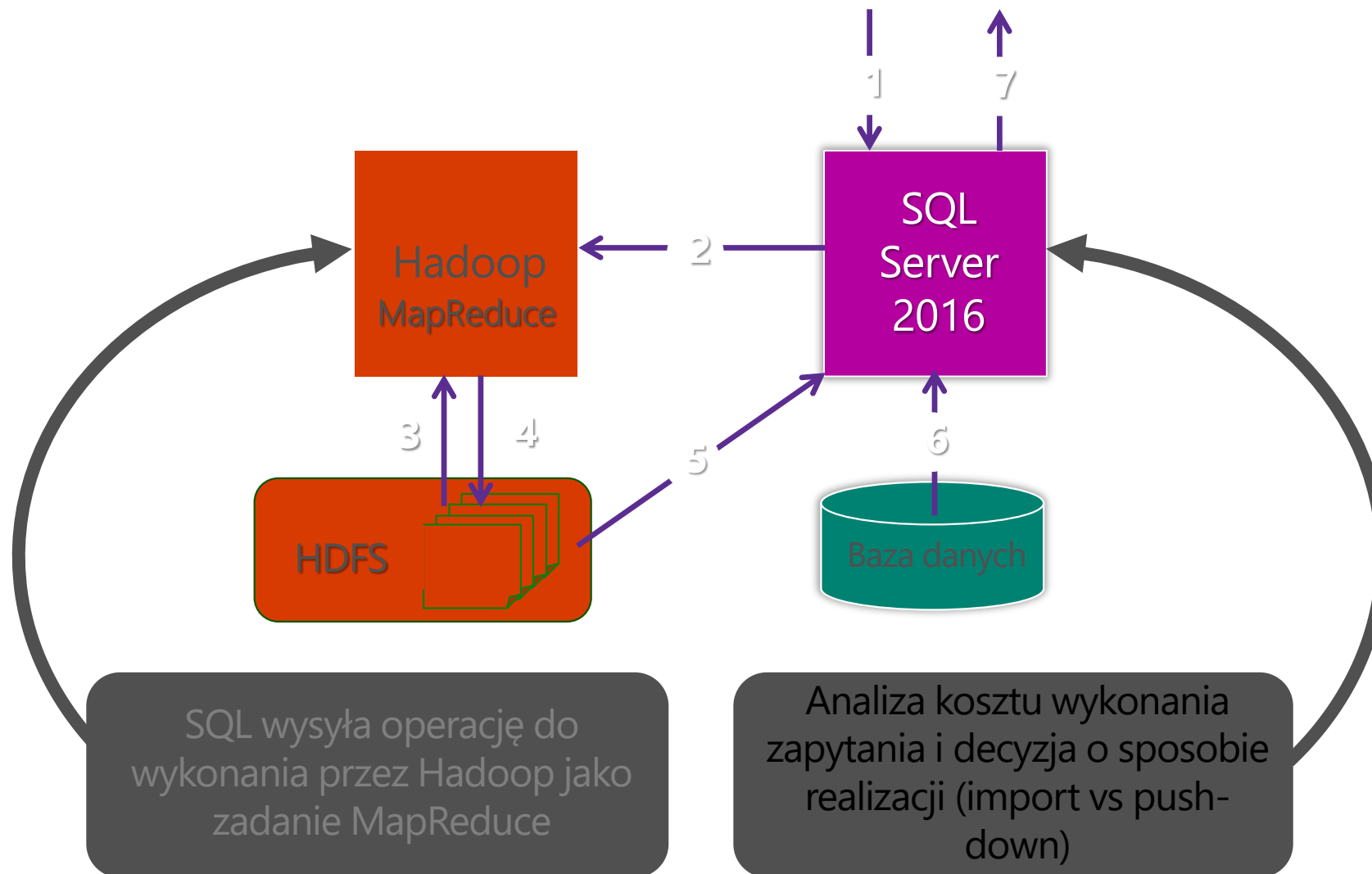
Dodatkowe widoki DMV :

[sys.dm_exec_compute_nodes](#)
[sys.dm_exec_compute_node_status](#)

-- usunięcie węzła z grupy
EXEC sp_polybase_leave_group;



Jak działa Polybase ?





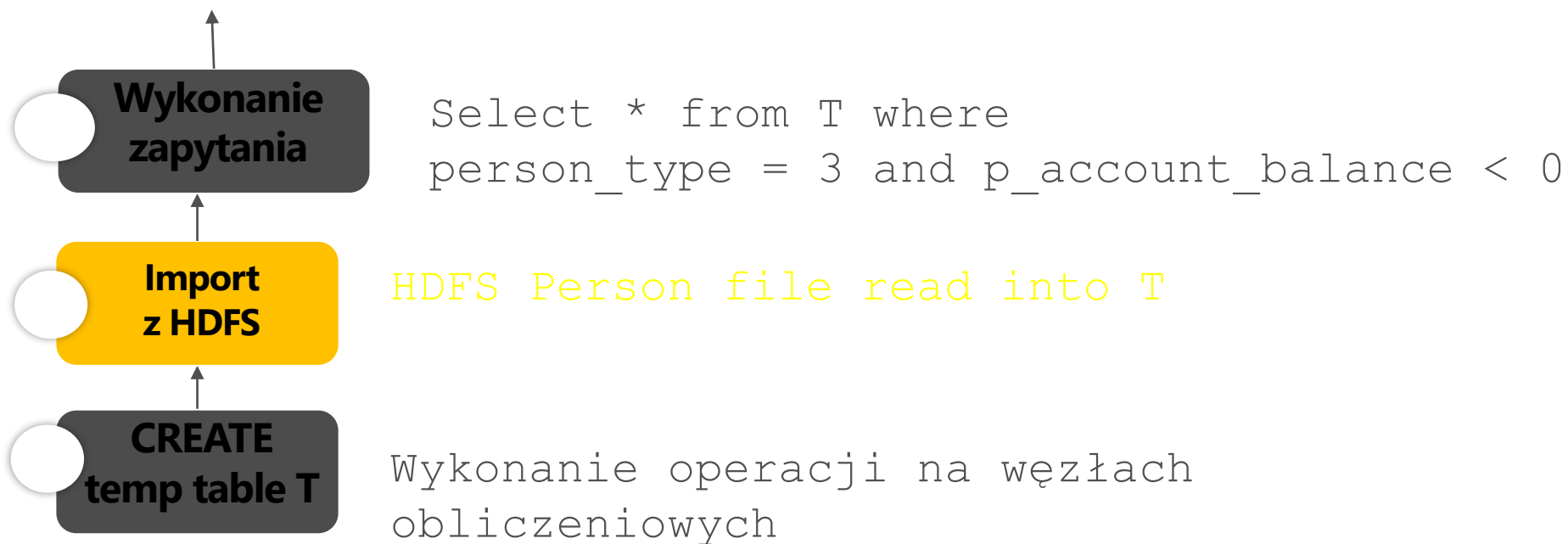
PolyBase przykładowe zapytanie

-- select on external table (data in HDFS)

```
SELECT * FROM Person
```

```
WHERE person_type = 3 and p_account_balance < 0;
```

Potencjalny plan wykonania zapytania:





PolyBase przykładowe zapytanie

-- select and aggregate on external table (data in HDFS)

```
SELECT AVG(p_account_balance) FROM Person
```

```
WHERE person_type = 3 and p_account_balance < 0;
```

Plan wykonania:

Co dzieje się w tym kroku ?

Krok 1: Kompilacja predykatu do Java i wygenerowanie zadania MapReduce (MR)

Krok 2: Silnik przekazuje zadane MR do klastra Hadoop. Wynik zwracany w postaci "hdfsTemp"



Zastosowanie filtra i
wyliczenie agregatu dla osoby



PolyBase przykładowe zapytanie

-- select and aggregate on external table (data in HDFS)

```
SELECT AVG(p_account_balance) FROM Person  
WHERE person_type = 3 and p_account_balance < 0;
```

Plan wykonania:



1. Predykaty i agregacje "wypychane" do klastra Hadoop jako zadania MapReduce
2. Optymalizator na podstawie kosztu wykonania zapytania decyduje o "wypchnięciu" zapytania

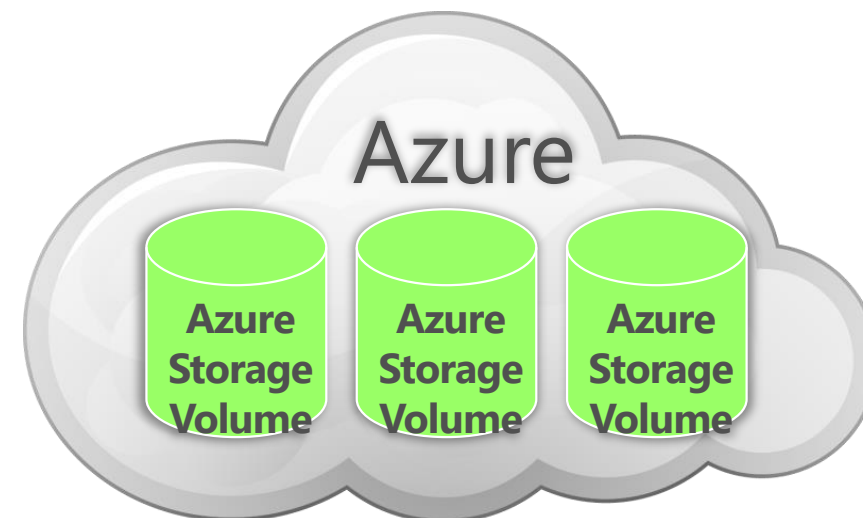
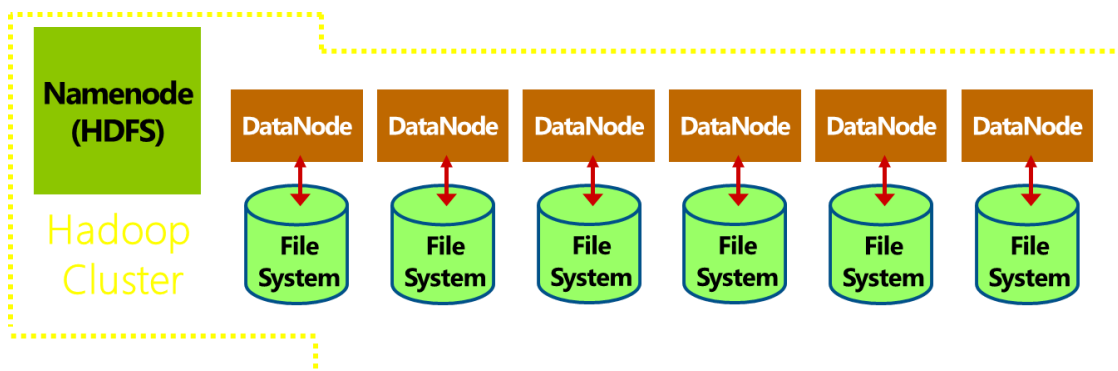


Jak zacząć z Polybase
oraz
SQL Server 2016?





Krok 1: Dostęp/konfiguracja do klastra Hadoop lub Azure Blob Storage



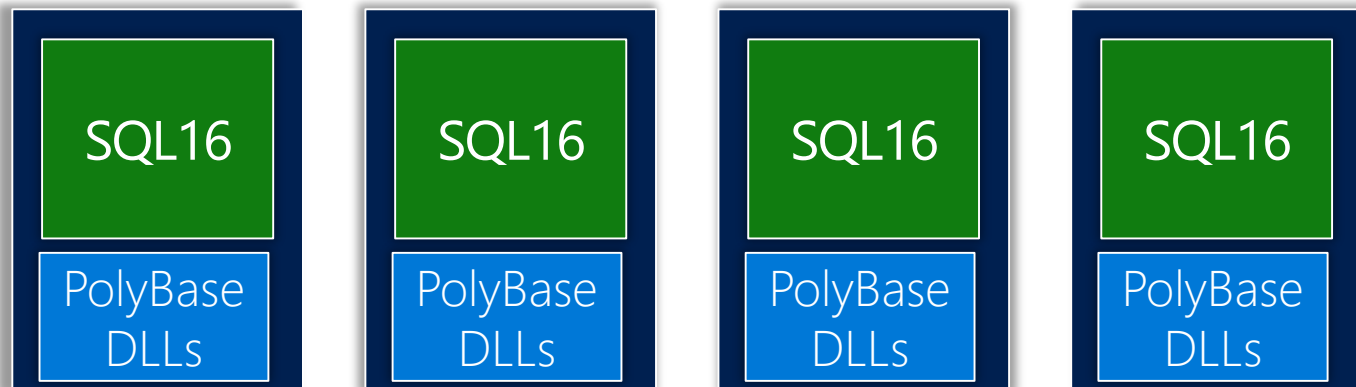
Hortonworks lub Cloudera

- Hadoop 2.0 lub wyższa
- Linux lub Windows
- On premises lub w Azure

- Azure Storage Blobs (ASB)
- PolyBase odczytuje/zapisuje dane na ASB korzystając z Hadoop RecordReaders/Writers
- Brak wsparcia dla compute push-down



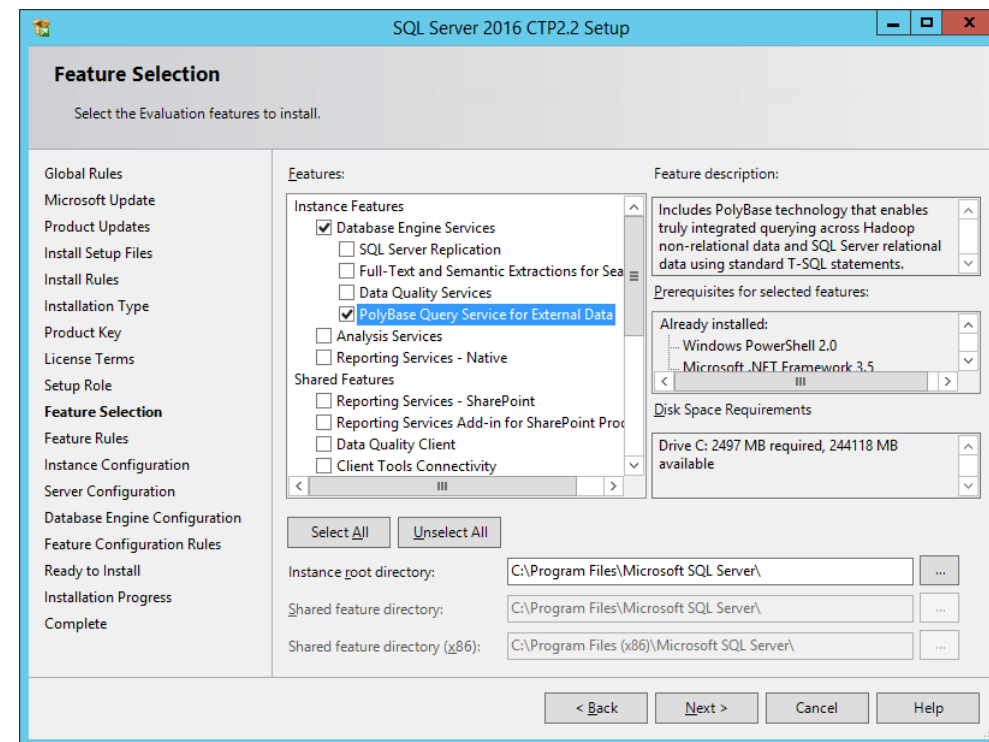
Krok 2: Instalacja SQL Server



Instalacja jednej lub kilku instancji SQL Server z obsługą PolyBase

PolyBase DLLs (Engine oraz DMS) są instalowane i rejestrowane jako usługi Windows

Wymagane: JRE (Oracle)





Określenie źródła danych Hadoop sp_configure

```
-- Run sp_configure 'hadoop connectivity'
-- and set an appropriate value
sp_configure
    @configname = 'hadoop connectivity',
    @configvalue = 7;
GO
RECONFIGURE
GO

-- List the configuration settings for
-- one configuration name
sp_configure @configname='hadoop connectivity';
GO
```

Dostępne wartości

- 0: Obsługa Hadoop wyłączona
- 1: Hortonworks HDP 1.3 on Windows Server
Azure blob storage (WASB[S])
- 2: Hortonworks HDP 1.3 on Linux
- 3: Cloudera CDH 4.3 on Linux
- 4: Hortonworks HDP 2.0 on Windows Server
Azure blob storage (WASB[S])
- 5: Hortonworks HDP 2.0 on Linux
- 6: Cloudera 5.1 on Linux
- 7: Hortonworks 2.1 and 2.2 on Linux
Hortonworks 2.2 on Windows Server
Azure blob storage (WASB[S])



Demo





Microsoft

Bartłomiej Graczyk
Data Platform Solution Architect
bagra@microsoft.com | bartek@graczyk.info.pl