Moduł 4
Technologie gromadzenia, przechowywania i zarządzania Big Data

iBigWorld:
Innovations for Big Data in a Real World

Zespół UBB

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Struktura – Tematy wykładów

- Wiedza podstawowa studentów dotycząca tematu wykładu;
- a. Główne cechy big data;
- b. Architektury big data.

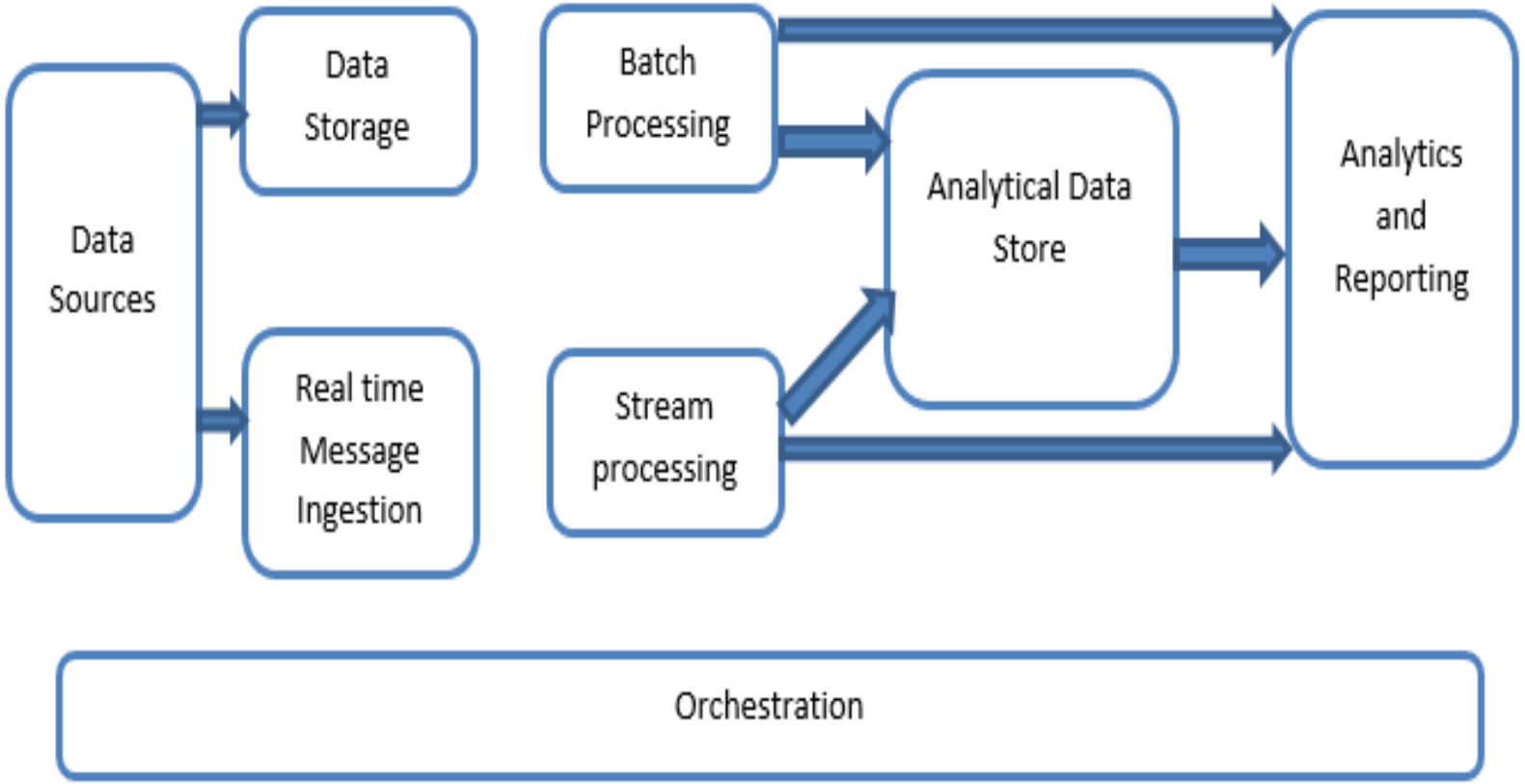
1. Technologie Wprowadzenie;
2. Ewolucja technologii przechowywania Big Data;
3. Taksonomia technologii przechowywania Big Data;
4. Analiza technologii przechowywania Big Data;
5. Wyzwania badawcze na przyszłość;
6. Wnioski.

a. Najważniejsze cechy Big data

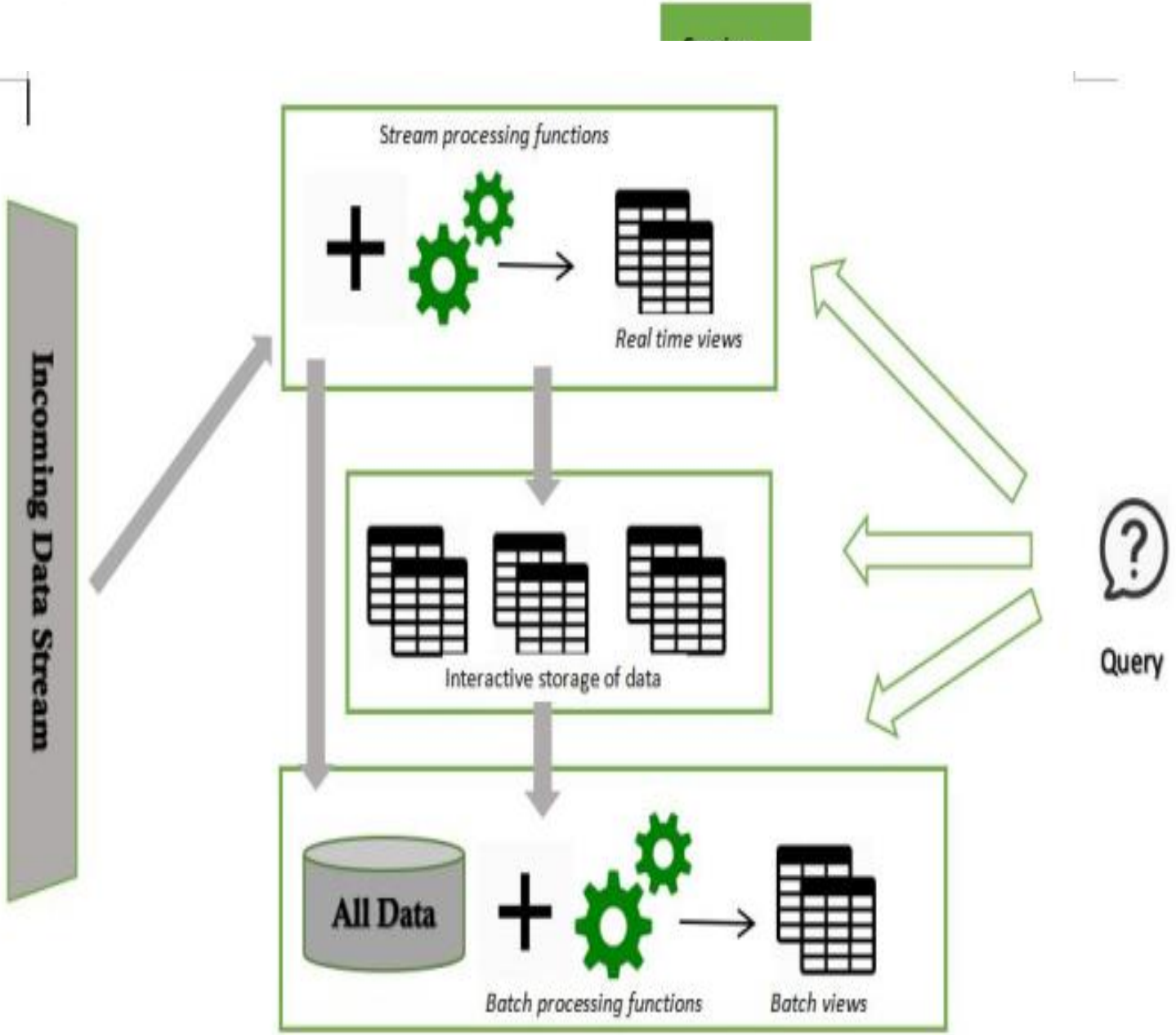
- Skalowalność;
- Dostępność;
- Odporność na błędy;
- Spójność;
- Indeksowanie wtórne.

b. Architektura Big Data

- Ogólne przetwarzanie i technologie systemu Bigdata:



- A. Architektura Lambda
- B. Architektura Kappa
- C. Architektura Microservice
- D. Architektura Zeta
- E. Architektura IoT



Na temat wykładu

- *Technologie Wprowadzenie;*
- *Ewolucja technologii przechowywania dużych danych;*
- *Taksonomia technologii przechowywania dużych danych;*
- *Analiza technologii przechowywania dużych danych;*
- *Wyzwania badawcze na przyszłość;*
- *Wnioski*

Kluczowe cele tego wykładu:

- (1) przedstawienie struktur pamięci masowej szerokiego zakresu technologii w środowisku big data;
- (2) podkreślić charakterystyczne właściwości każdej z technologii składowania;
- (3) rozwinąć taksonomię i ocenić technologie składowania big data zgodnie ze znanym twierdzeniem Brewera przedstawionym dla systemów rozproszonych;
- (4) określenie wyzwań i kierunków radzenia sobie z przechowywaniem big data w przyszłości.

Wykład jest zorganizowany w następujący sposób:

- **Część 2** opisuje ewolucję technologii przechowywania big data i ich cechy wyróżniające w stosunku do relacyjnych baz danych. Współczesne technologie przechowywania big data są również szczegółowo opisane w części 2.
- **Część 3** przedstawia taksonomię i kategoryzację w oparciu o przyjęty model danych i licencjonowanie.
- **Część 4** opisuje twierdzenie Brewera CAP dla systemów rozproszonych wraz z jego nowym wyjaśnieniem. Przedstawiono i przeanalizowano technologie pamięci masowych w celu zaproponowania ich rodzaju na podstawie kategoryzacji Brewera.
- **Część 5** podsumowuje dyskusję i wskazuje na przyszłe wyzwania badawcze. **Część 6** kończy dyskusję.

2020-1-PL01-KA203-082197

Część 2. Ewolucja technologii przechowywania dużych danych

- 2.1. Wyróżniki technologii przechowywania big data
- Analiza SWOT relacyjnych baz danych i systemów przechowywania big data:
- Mocne strony
- Słabe strony
- Szanse
- Zagrożenia

- 2.2 Współczesne technologie przechowywania

Erasmus+

2.2 Współczesne technologie przechowywania big data

- The Google File System (GFS, 2003);
- The Hadoop Distributed File System (HDFS, 2008);
- BigTable (2008);
- Hbase (2015);
- Hypertable (2015);
- MongoDB (2015);
- Terrastore;
- HyperGraphDB (2010);
- InfiniteGraph (2014);
- Rocket U2 (2015);

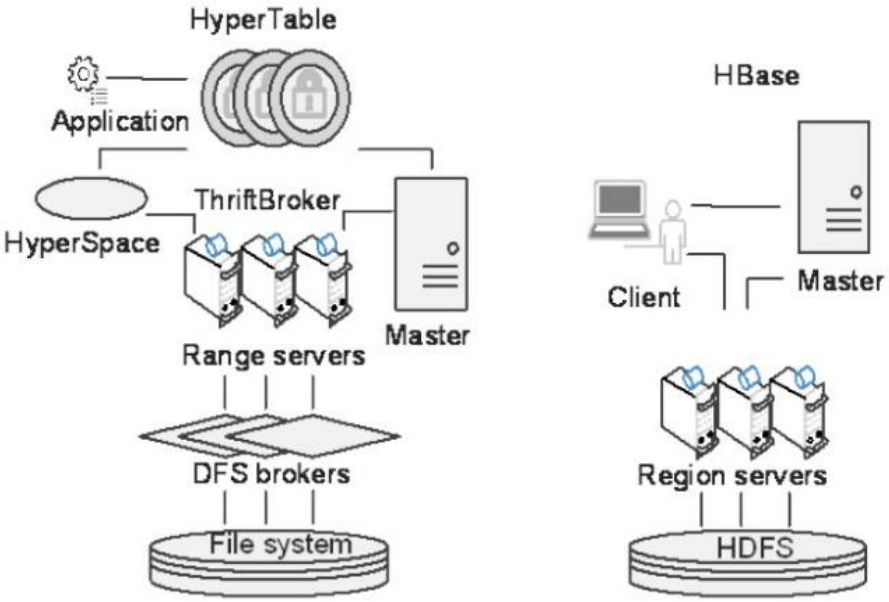


Fig. 3 Two implementations of BigTable (George, 2011; MongoDB, 2015)

2.2 Współczesne technologie przechowywania big data (2)

- Scalaris (2008);
- Berkeley DB (2015);
- DynamoDB (2012);
- Qizx (2014);
- Neo4j (2015);
- RethinkDB (2015);
- Aerospike (2015);
- OrientDB (2015);
- AllegroGraph (2015);

2.2 Contemporary big data storage technologies (3)

- Redis (2013);
- Voldemort (2015);
- KAI (2015);
- Cassandra (2010);
- SimpleDB (2007);
- MemcacheDB (2011);
- CouchDB (2010);
- Riak (2010);
- SciDB (2009).

Część 3. Taksonomia technologii przechowywania dużych danych

Istnieją cztery rodzaje modeli danych:

- ***key-value;***
- ***zorientowany na kolumny***
- ***zorientowany na dokumenty;***
- ***i graf.***

Natomiast licencjonowanie ma trzy kategorie:

- ***open source,***
- ***własnościowe,***
- ***i komercyjne.***

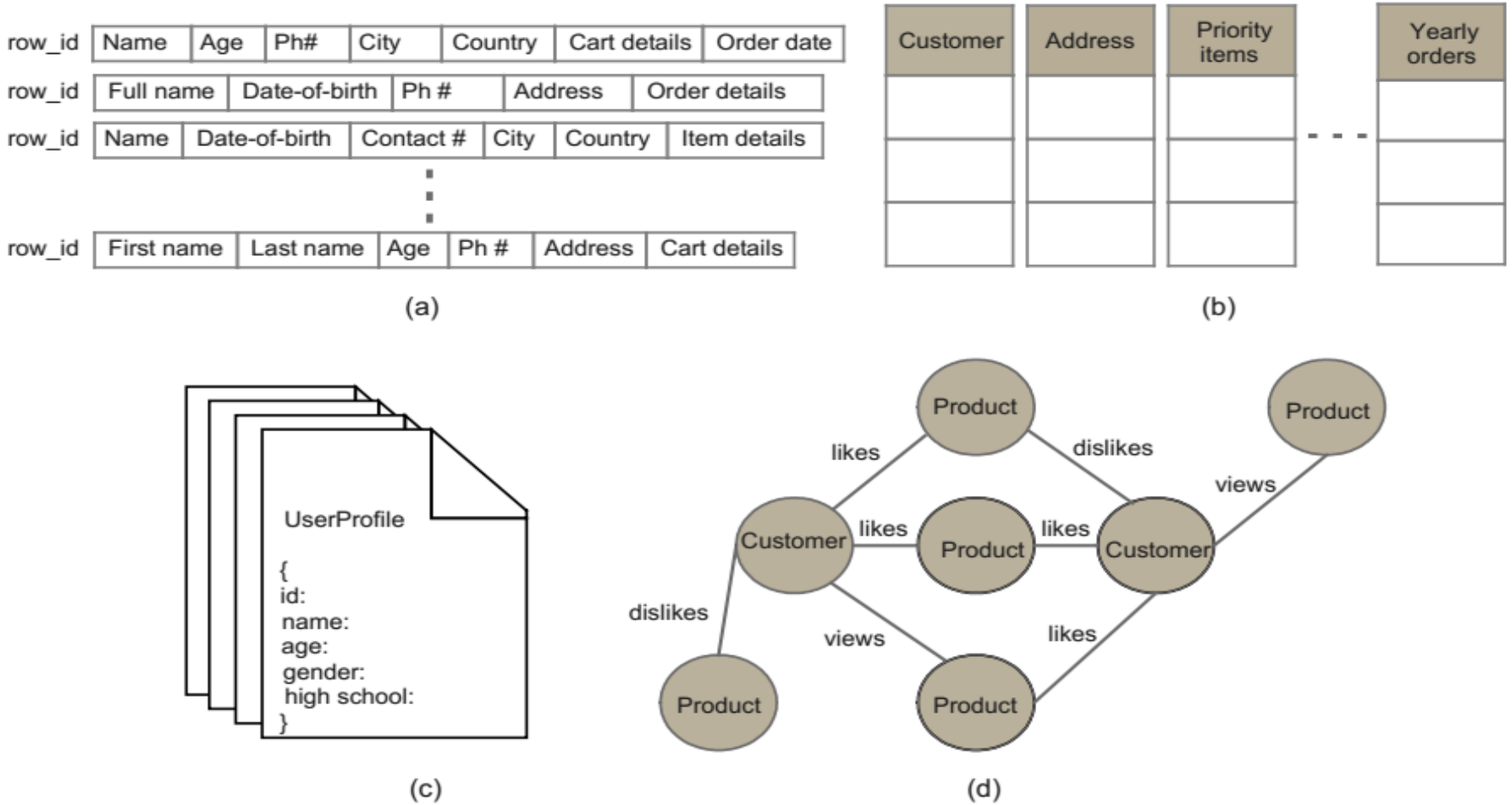


Fig. 4 Examples of data models: (a) key-value; (b) column-oriented; (c) document-oriented; (d) graph

Modele danych NoSQL

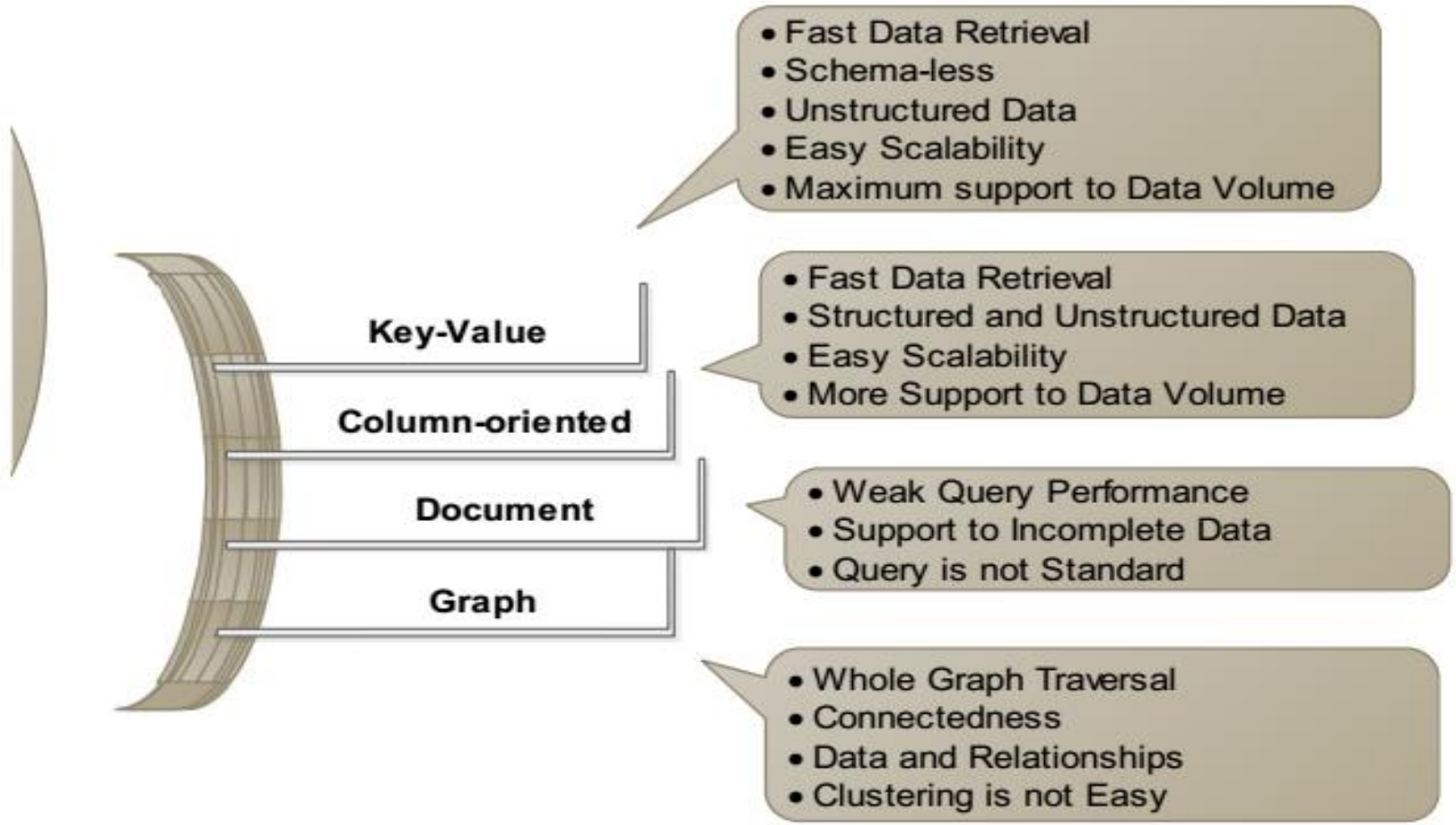


Fig. 5 Data models of NoSQL databases

Podsumowanie technologii przechowywania big data -

Technologia	Dostawca	Cele projektowe
BigTable	Google	Wprowadzenie dystrybucji dla wysoce skalowalnych, ustrukturyzowanych danych
HBase	Apache	Aby zapewnić spójny, losowy i w czasie rzeczywistym dostęp do skalowalnych. BigTables z żądaniami odczytu/zapisu
Hypertable	Zvents	Zapewnienie równoległych, wysokowydajnych, skalowalnych baz danych dla dużych rozmiarów danych; wspieranie lepszej wydajności zapytań dla dużych rozmiarów danych
MongoDB	MongoDB, Inc.	Provide parallel, high-performance, scalable databases for large data sizes; support improved query performance for large data sizes
Terrastore (in-memory)	Terracotta, Inc.	Osiągnięcie spójności dla danych dokumentów poprzez dystrybucję
HyperGraphDB	Kobrix software, Inc.	Zaprojektowanie modelu pamięci trwałej dla projektów z zakresu sztucznej inteligencji iprojektów sieci semantycznej; zapewnienie zarówno relacyjnego, jak i obiektowego zarządzania danymi
InfiniteGraph	Objectivity, Inc	Zapewnienie trwałego, rozproszonego przechowywania danych z łatwiejszym przemierzaniem złożonych relacji; wspieranie złożonych zapytań do danych w celu uzyskać wyższe wartości

Comparison and application areas of storage technologies – part of Table 3

Technology	W pamięci	Na dysku	Wytrwałość	Interaktywne pisanie/czytanie	Podział danych	Udostępnienie	Skalowalność	Zastosowanie
Scalaris	√	×	×	√	√	√	√	Skalowalne usługi online (np. eBay, Amazon); Bazy danych typu "always live"; Węzły o częstych awariach;
Aerospike	√	×	×	√	√	√	√	Aplikacje na skalę internetową; Dyski SSD
Redis	√	×	√	√	√	√	×	Pamięć podręczna oparta na sesji; Strukturalne ciągi znaków; LRU cache; Dla małych danych
Voldemort	×	√	√	√	√	×	√	Dane tylko do odczytu; LinkedIn;
KAI	×	√	√	–	√	–	×	Repozytorium internetowe; Sieć społecznościowa; Sklepy internetowe;
MemcacheDB	√	√	√	√	√	–	√	Przechowywanie obiektów; Dane tekstowe tylko do odczytu;

Opis udzielania licencji przedstawia się następująco :

- **Open source:** System open-source jest swobodnie dostępny zarówno dla środowiska akademickiego, jak i biznesu, dzięki czemu można go wykorzystać i zintegrować z własnym kawałkiem kodu lub aplikacją. Systemy te są opłacalne pod względem rozwoju i zapewniają lepszą jakość i elastyczny dostęp nawet dla małych firm.
- **Proprietary:** W przeciwieństwie do systemów open-source, systemy własnościowe można posiadać po uiszczeniu rozsądnej opłaty za ich użytkowanie. Systemy te są podawane z pewnymi warunkami prawnymi, aby uważać i tylko na własny użytek i nie być modyfikowanym lub redystrybuowanym. Kody źródłowe takich systemów zazwyczaj nie są przekazywane nabywcy.
- **Komercyjne:** Systemy te są tworzone na sprzedaż. Część próbna może być dostępna bezpłatnie, ale aby uzyskać pełne uprawnienia do badań i rozwoju; użytkownik lub przedsiębiorstwo musi je kupić.

Taksonomia technologii BigData;

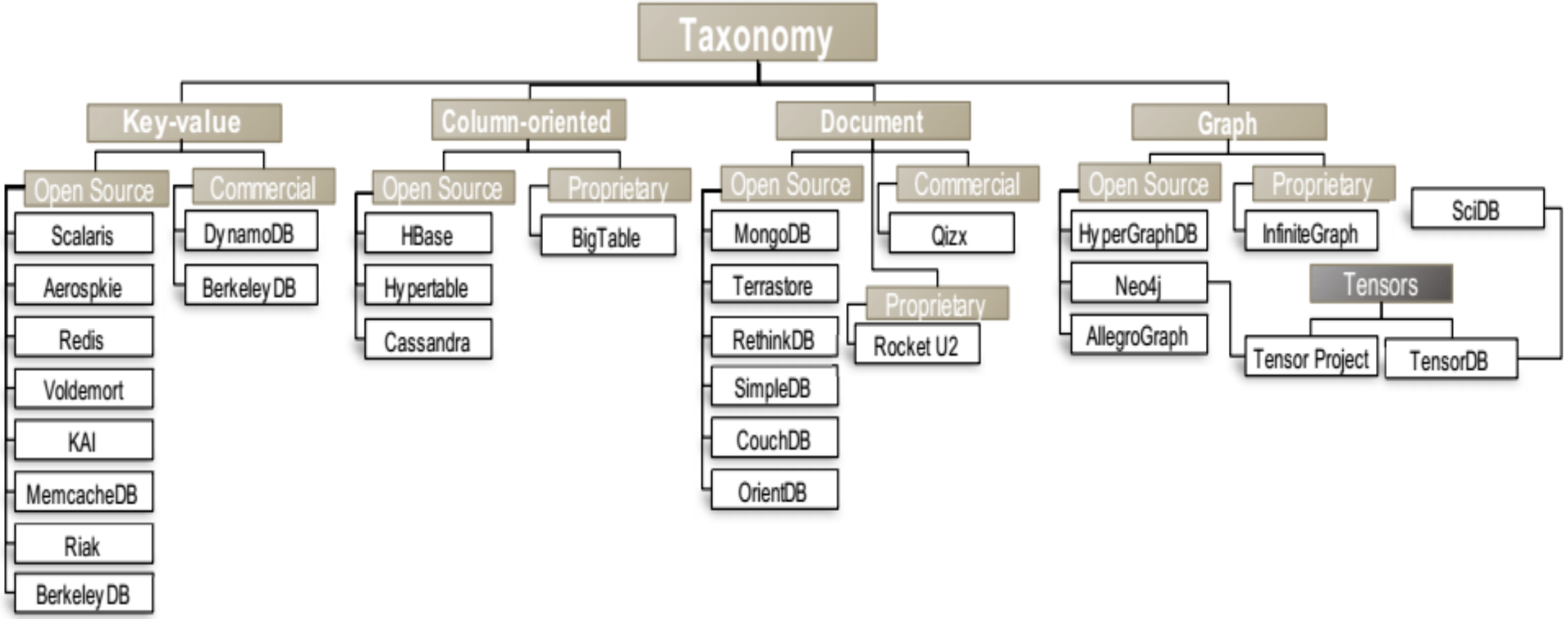


Fig. 6 Taxonomy of big data storage technologies

Część 4. Analiza technologii przechowywania big data

- *4.1. Twierdzenie CAP Brewera*
- *Twierdzenie CAP Brewera - podające powszechnie stosowane kryteria klasyfikacji systemów rozproszonych, stwierdza, że technologiom NoSQL trudno jest spełnić kryteria ACID lub BASE, ponieważ spójność, dostępność i odporność na partycje są istotnymi czynnikami przy projektowaniu tych wielkoskalowych, rozproszonych systemów pamięci masowych.*
- *ACID - Atomowość, Spójność, Izolacja, Trwałość.*
- *BASE - Basically Available, Soft State, Eventual Consistency .*

- *4.2. Analiza technologii przechowywania big data w oparciu o twierdzenie CAP Brewera*

4.1. Twierdzenie CAP Brewera

- Zasada CAP:
- Spójność odnosi się do posiadania tych samych, aktualnych danych w każdym węźle,
- Dostępność sugeruje szybki dostęp do zasobów przechowywania danych przy minimalnym czasie przestoju,
- oraz Odporność partycji związana jest z odpornością na błędy w przypadku braku reakcji węzłów lub podsieci.

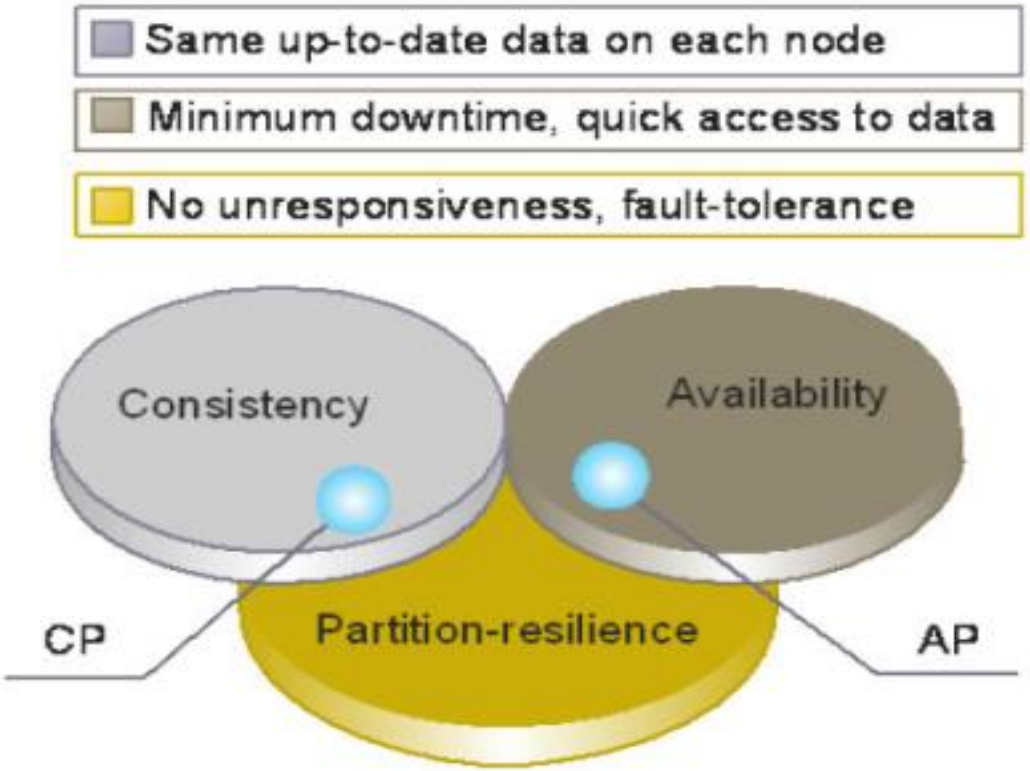


Fig. 7 Brewer's CAP theorem (Brewer, 2012)

4.2. Analiza technologii przechowywania big data w oparciu o twierdzenie CAP Brewera

- Oczywiste jest, że system pamięci masowej będzie albo:
- typu CP (consistency-partition resilience)
- lub typu AP (availability-partition resilience).
- **Tabela 5** - kategoryzuje system na typy CP i AP, ponieważ w twierdzeniu CAP zadeklarowano, że niemożliwe jest, aby baza danych posiadała jednocześnie spójność ACID i dostępność danych.
- **W tabeli 5** opracowano cechy technologii przechowywania big data:
- Cechy te są dalej brane pod uwagę przy analizie i kategoryzacji tych systemów przechowywania danych zgodnie z twierdzeniem Brewera.
- **Tabela 5** opisuje również metody implementacji spójności, replikacji, partycjonowania danych i indeksowania dla tych technologii.

Part of Table 5

Dane model	Licencja	Tchnologia	Dane przechowywane(St)/ Dane bazowe(B)	Cechy	Język zapytań	Konsystencja	Replikacja	Podział	Indeksowania	Brewer's category
		Scalaris	St	Silna spójność Skalowalność i wysoka dostępność dzięki równoważeniu obciążenia i odporności na błędy Bardzo małe koszty utrzymania Samodzielne zarządzanie	Custom	E	Symm	K	Pr	CP
		Aerospike	St	Wysoka skalowalność, spójność i niezawodność	AQL	St	Syn, Asyn	Sh	Sc	AP
		Redis	St	Automatyczne dzielenie na partycje Wydajny dostęp do odczytu/zapisu danych	-	E	MS	CH	C	CP

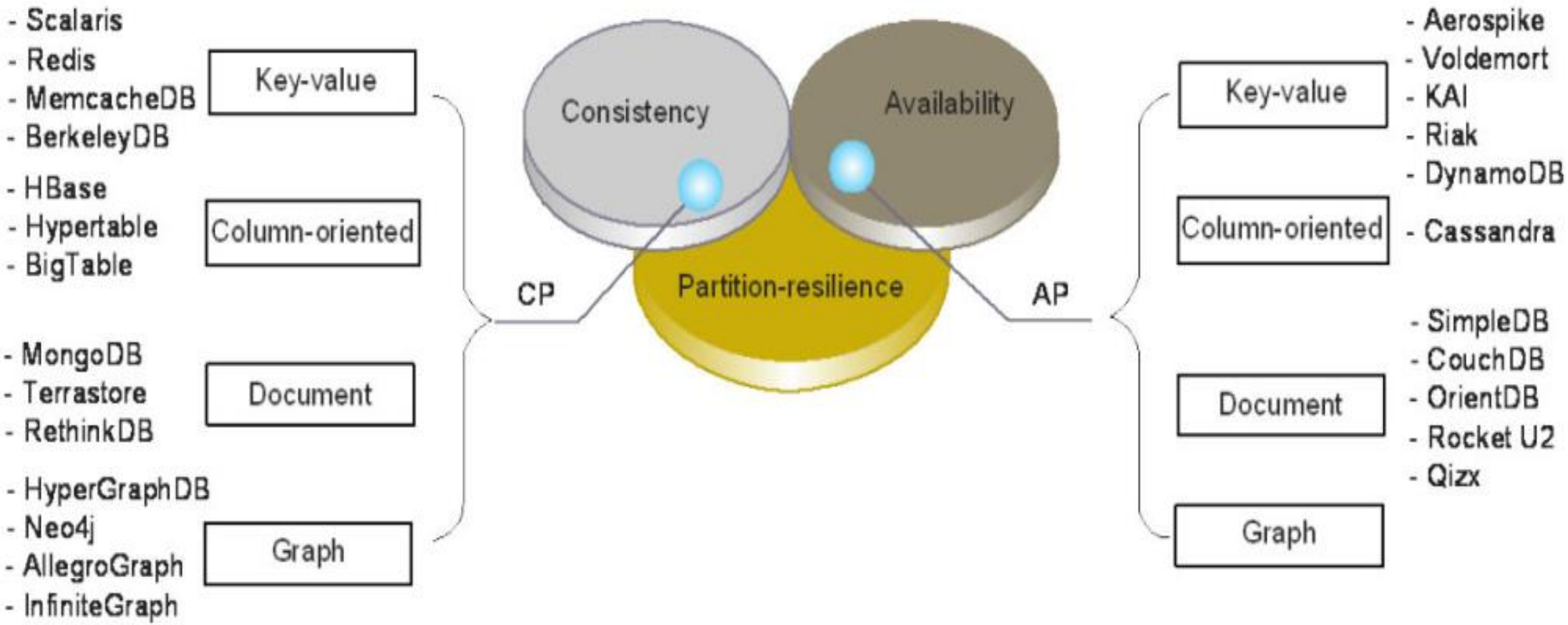


Fig. 8 Big data storage technologies and the CAP theorem

5. Wyzwania związane z przyszłymi badaniami

- *1. Częsta aktualizacja danych i zmiana schematu:*
- *2. Metoda partycjonowania:*
- *3. Współczynnik replikacji:*
- *4. Wiedza użytkownika:*

6. Podsumowanie

- ***Koniec***