Big Data?

Agenda

1. Czym jest Big Data?
2. Cele i historia powstania
3. Typowe zastosowania
4. Big Data a bazy relacyjne
5. Big Data a Data Science

BIG DATA

- Velocity
- Variety
- Value
- Volume
- Veracity

BIG DATA

- Velocity 
- Variety
- Value
- Volume
- Veracity

BIG DATA

- Velocity
- Variety
- Value
- Volume
- Veracity

BIG DATA

- Velocity 
- Variety 
- Value 
- Volume
- Veracity

BIG DATA

- Velocity
- Variety
- Value
- Volume
- Veracity

BIG DATA

- Velocity 
- Variety 
- Value 
- Volume 
- Veracity 

Historia w pigułce

10.2002

Powstaje projekt Nutch

Na rynku pojawia się Intel Celeron 1.3GHz

Historia w pigułce

10.2002	10.2003
Powstaje projekt Nutch	**Google File System white paper**
Na rynku pojawia się Intel Celeron 1.3GHz	Powstaje standard Serial ATA

Historia w pigułce

10.2002	10.2003	10.2004
Powstaje projekt Nutch	**Google File System white paper**	**Google Map Reduce white paper**
Na rynku pojawia się Intel Celeron 1.3GHz	Powstaje standard Serial ATA	

Historia w pigułce

10.2002	10.2003	10.2004	01.2006
Powstaje projekt Nutch	**Google File System white paper**	**Google Map Reduce white paper**	**Powstaje projekt Hadoop**
Na rynku pojawia się Intel Celeron 1.3GHz	Powstaje standard Serial ATA		Pierwszy dysk 750GB. Pojawia się procesor Intel Core 2 Duo 1.86GHz

Historia w pigułce

10.2002	10.2003	10.2004	01.2006	04.2006
Powstaje projekt Nutch	**Google File System white paper**	**Google Map Reduce white paper**	**Powstaje projekt Hadoop**	**Pierwsza wersja Hadoopa - 0.1.0**
Na rynku pojawia się Intel Celeron 1.3GHz	Powstaje standard Serial ATA		Pierwszy dysk 750GB. Pojawia się procesor Intel Core 2 Duo 1.86GHz	Udało się posortować 1.8TB danych - 189 nodów - 47.9 godzin

Historia w pigułce

10.2003	10.2004	01.2006	04.2006	04.2008
Google File System white paper	**Google Map Reduce white paper**	**Powstaje projekt Hadoop**	**Pierwsza wersja Hadoopa - 0.1.0**	**1TB Sort - 209 sekund - 910 nodów**
Powstaje standard Serial ATA		Pierwszy dysk 750GB. Pojawia się procesor Intel Core 2 Duo 1.86GHz	Udało się posortować 1.8TB danych - 189 nodów - 47.9 godzin	Pierwszy dysk 1TB. Na rynek wchodzi procesor o taktowaniu 2.8GHz

Historia w pigułce

10.2004	01.2006	04.2006	04.2008	06.2008
Google Map Reduce white paper	**Powstaje projekt Hadoop** Pierwszy dysk 750GB. Pojawia się procesor Intel Core 2 Duo 1.86GHz	**Pierwsza wersja Hadoopa - 0.1.0** Udało się posortować 1.8TB danych - 189 nodów - 47.9 godzin	**1TB Sort - 209 sekund - 910 nodów** Pierwszy dysk 1TB. Na rynek wchodzi procesor o taktowaniu 2.8GHz	**Powstaje projekt Hive**

Historia w pigułce

01.2006	04.2006	04.2008	06.2008	08.2008
Powstaje projekt Hadoop	**Pierwsza wersja Hadoopa - 0.1.0**	**1TB Sort - 209 sekund - 910 nodów**	**Powstaje projekt Hive**	**Pierwsza firma zajmująca się Hadoopem**
Pierwszy dysk 750GB. Pojawia się procesor Intel Core 2 Duo 1.86GHz	Udało się posortować 1.8TB danych - 189 nodów - 47.9 godzin	Pierwszy dysk 1TB. Na rynek wchodzi procesor o taktowaniu 2.8GHz		Cloudera.

Historia w pigułce

04.2006	04.2008	06.2008	08.2008	05.2009
Pierwsza wersja Hadoopa - 0.1.0	**1TB Sort - 209 sekund - 910 nodów**	**Powstaje projekt Hive**	**Pierwsza firma zajmująca się Hadoopem**	**1TB Sort - 62 sekundy**
Udało się posortować 1.8TB danych - 189 nodów - 47.9 godzin	Pierwszy dysk 1TB. Na rynek wchodzi procesor o taktowaniu 2.8GHz		Cloudera.	Umiemy też sortować 1PB. Na rynku AMD Athlon quad 3.1GHz

Historia w pigułce

04.2008	06.2008	08.2008	05.2009	06.2011
1TB Sort - 209 sekund - 910 nodów	**Powstaje projekt Hive**	**Pierwsza firma zajmująca się Hadoopem**	**1TB Sort - 62 sekundy**	**Druga firma zajmująca się biznesowo Hadoopem**
Pierwszy dysk 1TB. Na rynek wchodzi procesor o taktowaniu 2.8GHz		Cloudera.	Umiemy też sortować 1PB. Na rynku AMD Athlon quad 3.1GHz	Hortonworks

Historia w pigułce

06.2008	08.2008	05.2009	06.2011	12.2011
Powstaje projekt Hive	**Pierwsza firma zajmująca się Hadoopem** Cloudera.	**1TB Sort - 62 sekundy** Umiemy też sortować 1PB. Na rynku AMD Athlon quad 3.1GHz	**Druga firma zajmująca się biznesowo Hadoopem** Hortonworks	**Hadoop 1.0** HDFS dostaje tryb HA

Historia w pigułce

08.2008	05.2009	06.2011	12.2011	08.2012
Pierwsza firma zajmująca się Hadoopem	**1TB Sort - 62 sekundy**	**Druga firma zajmująca się biznesowo Hadoopem**	**Hadoop 1.0**	**Powstaje YARN (MR v2)**
Cloudera.	Umiemy też sortować 1PB. Na rynku AMD Athlon quad 3.1GHz	Hortonworks	HDFS dostaje tryb HA	

Historia w pigułce

05.2009	06.2011	12.2011	08.2012	2012
1TB Sort - 62 sekundy Umiemy też sortować 1PB. Na rynku AMD Athlon quad 3.1GHz	**Druga firma zajmująca się biznesowo Hadoopem** Hortonworks	**Hadoop 1.0** HDFS dostaje tryb HA	**Powstaje YARN (MR v2)**	**Every minute of every day we create:** **- Over 2 million Google search queries** **- More than 100,000 tweets**

Historia w pigułce

06.2011	12.2011	08.2012	2012	10.2013
Druga firma zajmująca się biznesowo Hadoopem Hortonworks	**Hadoop 1.0** HDFS dostaje tryb HA	**Powstaje YARN (MR v2)**	**Every minute of every day we create:** **- Over 2 million Google search queries** **- More than 100,000 tweets**	**Hadoop 2.2** Pierwsza wersja linii 2. YARN + HDFS HA

Historia w pigułce

12.2011	08.2012	2012	10.2013	02.2014
Hadoop 1.0 HDFS dostaje tryb HA	**Powstaje YARN (MR v2)**	**Every minute of every day we create:** **- Over 2 million Google search queries** **- More than 100,000 tweets**	**Hadoop 2.2** Pierwsza wersja linii 2. YARN + HDFS HA	**Powstaje Apache Spark** Pierwszy dysk 8TB. Dyski 2TB dostępne na rynku.

Historia w pigułce

08.2012	2012	10.2013	02.2014	12.2017
Powstaje YARN (MR v2)	**Every minute of every day we create:** **- Over 2 million Google search queries** **- More than 100,000 tweets**	**Hadoop 2.2** Pierwsza wersja linii 2. YARN + HDFS HA	**Powstaje Apache Spark** Pierwszy dysk 8TB. Dyski 2TB dostępne na rynku.	**Hadoop 3.0** Pierwsze dyski 12TB i 14TB. Dyski 8TB dostępne na rynku.

Historia w pigułce

2012	10.2013	02.2014	12.2017	2017
Every minute of every day we create: **- Over 2 million Google search queries** **- More than 100,000 tweets**	**Hadoop 2.2** Pierwsza wersja linii 2. YARN + HDFS HA	**Powstaje Apache Spark** Pierwszy dysk 8TB. Dyski 2TB dostępne na rynku.	**Hadoop 3.0** Pierwsze dyski 12TB i 14TB. Dyski 8TB dostępne na rynku.	**Every minute we create: 456k tweets; 3.6M searches to google**

Zastosowania

- Rekomendacje, profilowanie, personalizacja

Zastosowania

- Rekomendacje, profilowanie, personalizacja
- Analiza danych z social media

Zastosowania

- Rekomendacje, profilowanie, personalizacja
- Analiza danych z social media
- Wykrywanie oszustw

Zastosowania

- Rekomendacje, profilowanie, personalizacja
- Analiza danych z social media
- Wykrywanie oszustw
- Bezpieczeństwo

Zastosowania

- Rekomendacje, profilowanie, personalizacja
- Analiza danych z social media
- Wykrywanie oszustw
- Bezpieczeństwo
- Opieka zdrowotna

Zastosowania

- Rekomendacje, profilowanie, personalizacja
- Analiza danych z social media
- Wykrywanie oszustw
- Bezpieczeństwo
- Opieka zdrowotna
- IoT

Zastosowania

- Rekomendacje, profilowanie, personalizacja
- Analiza danych z social media
- Wykrywanie oszustw
- Bezpieczeństwo
- Opieka zdrowotna
- IoT
- Optymalizacja procesów biznesowych

Zastosowania

- Rekomendacje, profilowanie, personalizacja
- Analiza danych z social media
- Wykrywanie oszustw
- Bezpieczeństwo
- Opieka zdrowotna
- IoT
- Optymalizacja procesów biznesowych
- Sport

Zastosowania

- Rekomendacje, profilowanie, personalizacja
- Analiza danych z social media
- Wykrywanie oszustw
- Bezpieczeństwo
- Opieka zdrowotna
- IoT
- Optymalizacja procesów biznesowych
- Sport
- Finanse, ryzyko kredytowe, high speed trading

Zastosowania

- Rekomendacje, profilowanie, personalizacja
- Analiza danych z social media
- Wykrywanie oszustw
- Bezpieczeństwo
- Opieka zdrowotna
- IoT
- Optymalizacja procesów biznesowych
- Sport
- Finanse, ryzyko kredytowe, high speed trading
- Nauka

Big Data vs bazy relacyjne

	RDBMS	Big Data
Typ danych	Ustrukturyzowane	Bez znaczenia
Przepustowość	Niska	Wysoka
Skalowanie	Wertykalne	Horyzontalne
Spójność danych	ACID	Brak
Rozmiar danych	setki gigabajtów	setki terabajtów/petabajtów
Metoda dostępu	SQL	Różne frameworki
Czas odpowiedzi	milisekundy+	sekundy+
Formaty danych	zależny od bazy	CSV, XML, Text etc.

Big Data vs Data Science