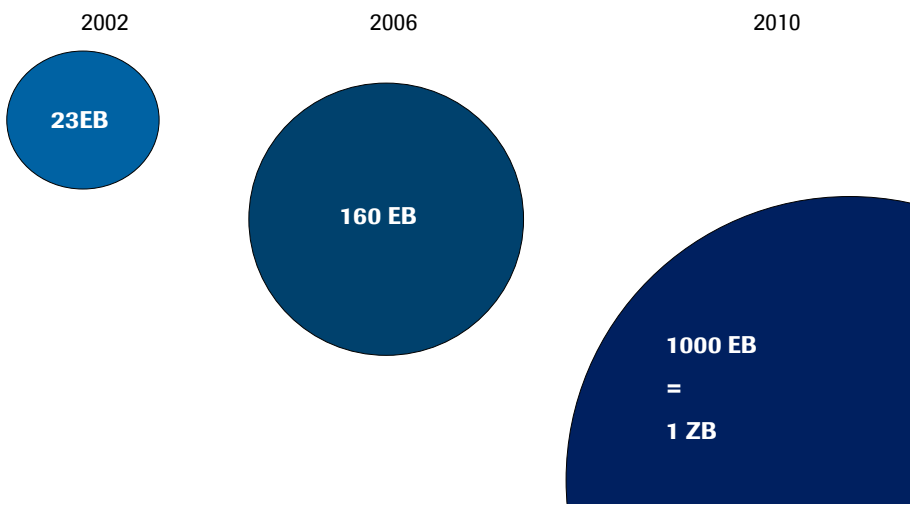
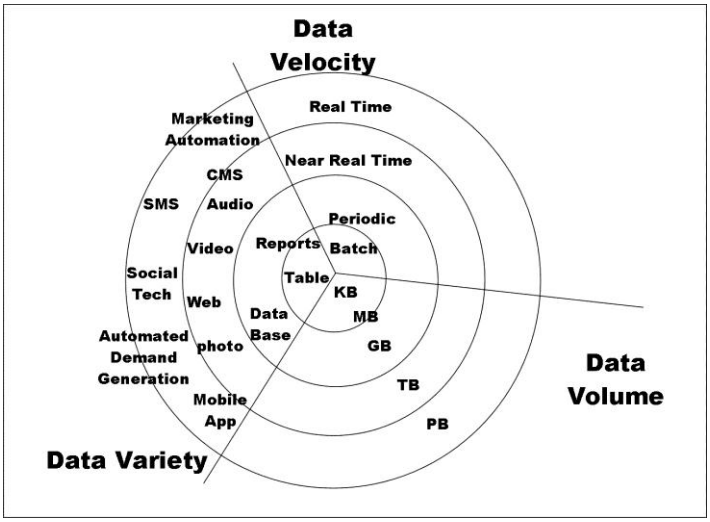
BIG DATA
2013-04-19

fabian wiktorowski

Historia



Definicja Big Data



http://jeffhurtblog.com/2012/07/20/three-vs-of-big-data-as-applied-conferences/

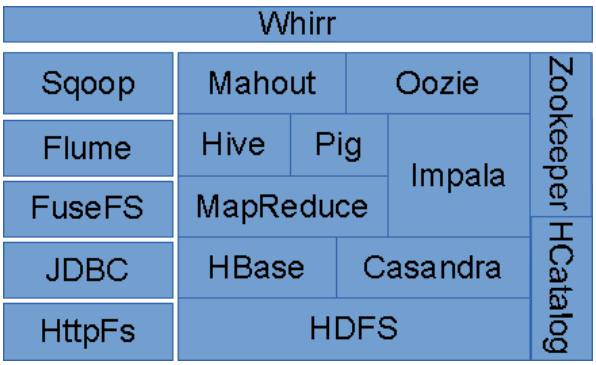
Inspiracja

- Google File System
- Google Big Table
- Map Reduce

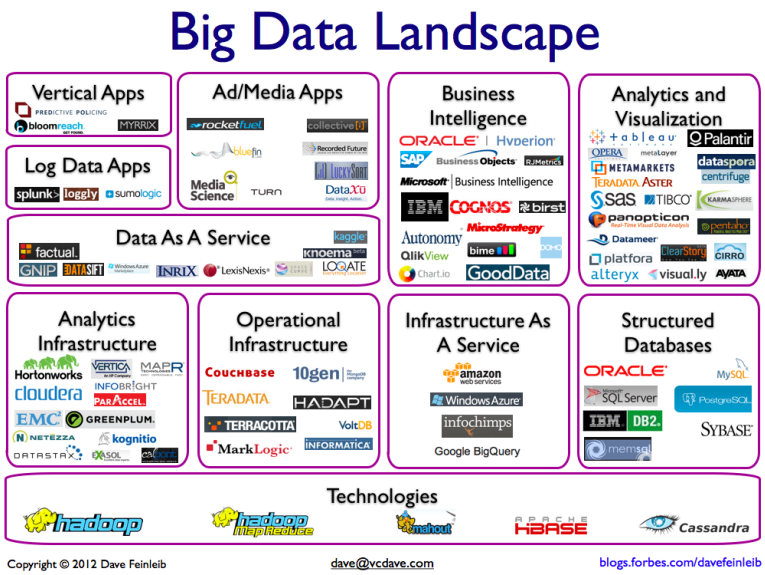
- Materiały:
  - http://research.google.com/archive/gfs.html
  - http://research.google.com/archive/bigtable.html
  - http://research.google.com/archive/mapreduce.html

Hadoop Ekosystem

- http://university.cloudera.com/onlineresources/hadoopecosystem.html

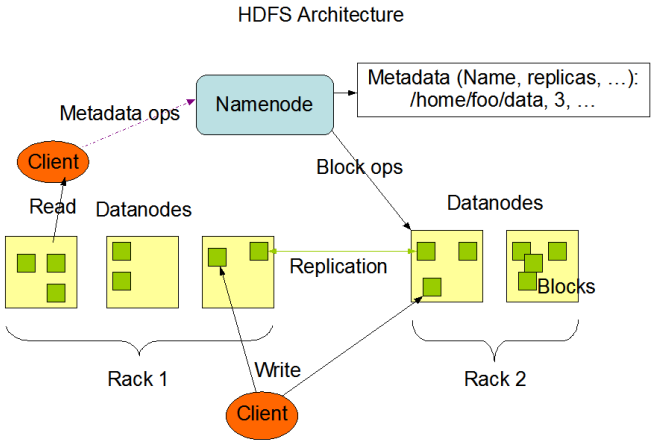


Dostawcy i technologie



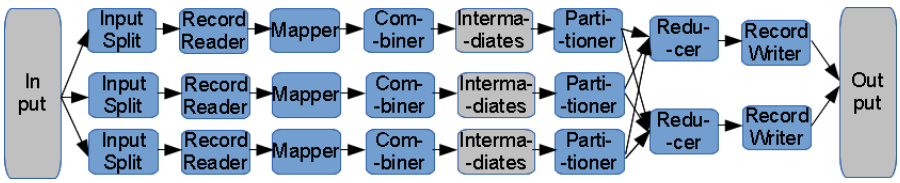
Hadoop HDFS

- Architektura http://hadoop.apache.org/docs/r1.0.4/hdfs_design.html



Map Reduce I

- Map
- Sort and Shuffle
- Reduce



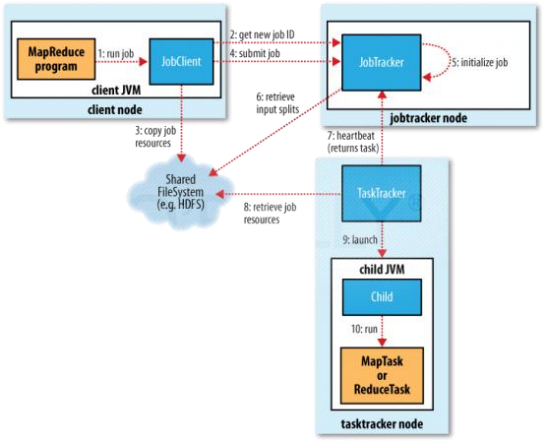
Map Reduce II

- Licencjonowanie
- Przykłady:
  - Zliczanie słów
  - Odwrócone indeksy
  - R http://www.bytemining.com/2010/08/taking-r-to-the-limit-part-ii-large-datasets-in-r/
  - Oracle https://blogs.oracle.com/datawarehousing/entry/in-database_map-reduce
- http://atbrox.com/2011/11/09/mapreduce-hadoop-algorithms-in-academic-papers-5th-update-%E2%80%93-nov-2011/

Wrappery i alternatywy dla Map Reduce

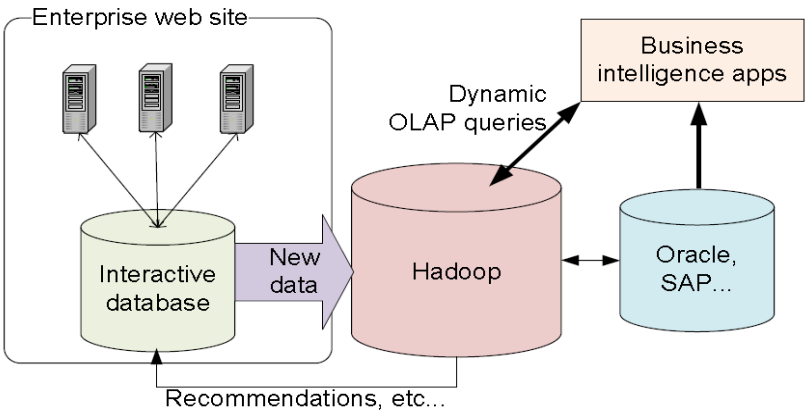
- Pig
- Hive
- Streaming
- Impala (http://blog.cloudera.com/blog/2012/10/cloudera-impala-real-time-queries-in-apache-hadoop-for-real/)
- Mahout
- Solr

Jak działa Map Reduce



- http://answers.oreilly.com/topic/459-anatomy-of-a-mapreduce-job-run-with-hadoop/
- http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

Hadoop w infrastrukturze



http://www.cloudera.com/content/cloudera/en/resources/library/recordedwebinar/webinar-integrating-hadoop-data-warehouse-business-intelligence-environment.html

Kolumnowa baza HBase

- Porównanie z bazami RDBMS

	RDBMS	HBase
Składowanie danych	Row-oriented	Column-Oriented
transakcyjność	tak	nie
język zapytań	sql	get/put/scan
indeksy	tak	row-key
ilość danych	terabajty	petabajty
przepustowość (na sekunde)	tysiące zapytań	miliony zapytań
zapisy	szybkie	powolne

Linki

- **DataSift** http://highscalability.com/blog/2011/11/29/datasift-architecture-realtime-datamining-at-120000-tweets-p.html
- **Hadoop at Yahoo!** http://developer.yahoo.com/blogs/ydn/posts/2013/02/hadoop-at-yahoo-more-than-ever-before/
- **hadoop** http://hadoopblog.blogspot.ch/
- **hadoop at Twitter** https://twitter.com/hadoop i http://engineering.twitter.com/search?q=hadoop
- **Cloudera** http://blog.cloudera.com/ i http://www.cloudera.com/content/cloudera/en/resources.html
- **Hadoop at Apache** http://hadoop.apache.org/

KSIĄŻKI

- **Hadoop: The Definitive Guide** http://www.amazon.com/Hadoop-Definitive-Guide-Tom-White/dp/1449311520/ref=dp_ob_title_bk
- **HBase: The Definitive Guide** http://www.amazon.com/HBase-Definitive-Guide-Lars-George/dp/1449396100/ref=pd_sim_b_1
- **Programming Pig** http://www.amazon.com/Programming-Pig-Alan-Gates/dp/1449302645/ref=pd_sim_b_1
- **Programming Hive** http://www.amazon.com/Programming-Hive-Edward-Capriolo/dp/1449319335/ref=pd_sim_b_1

VIDEO

- **Hadoop** http://www.youtube.com/watch?v=9s-vSeWej1U&NR=1&feature=fvwp
- **HBASE** (IBM) http://www.youtube.com/watch?v=XtLXPLb6EXs
- **Demystifying Hadoop** http://www.youtube.com/watch?v=xJHv5t8jcM8

Doing now what patients need next