Hadoop

Jakub Podeszwik

infoShare Academy

05.09.2016

Warsztat

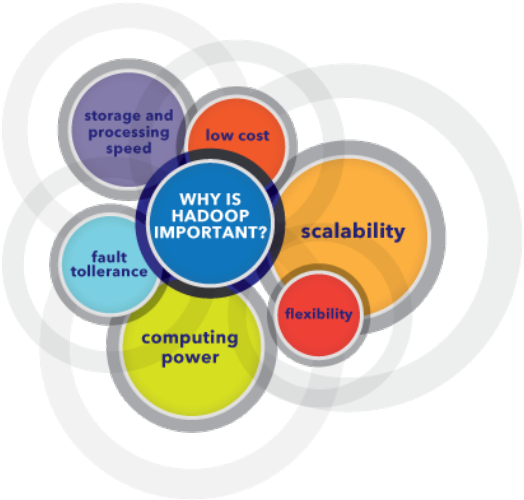
- 2 dni
- 20% prezentacja
- 40% live coding
- 40% zadań głównie programistycznych

Dzień 1

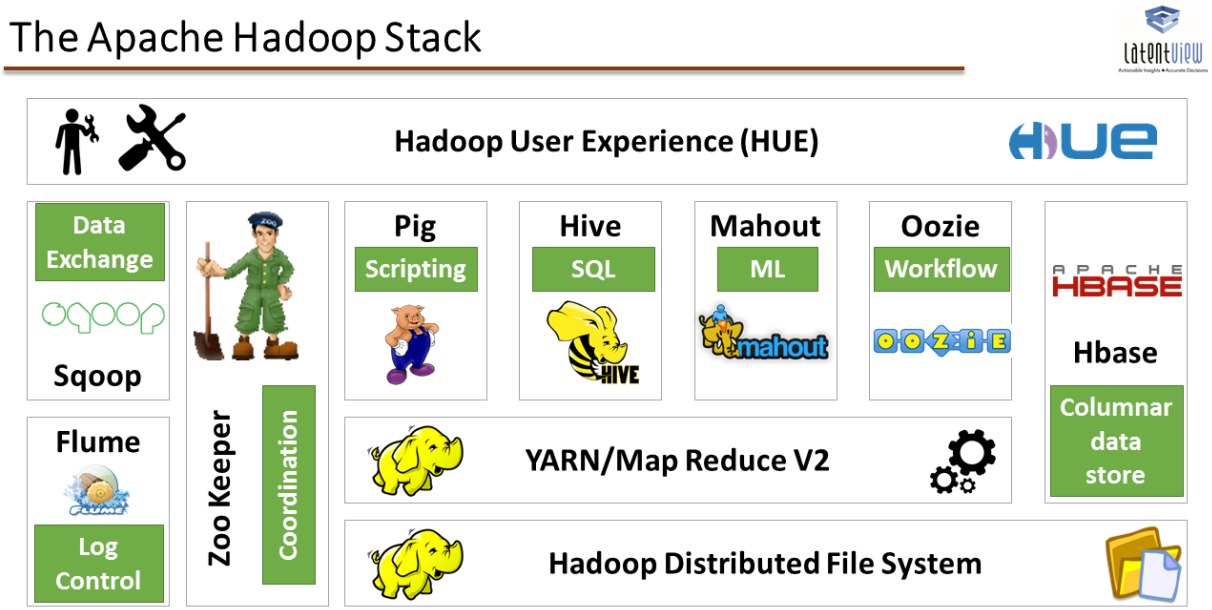
1. Wykład
    1. Czym jest hadoop?
    2. Komponenty
    3. Dystrybucje
2. Zajęcia praktyczne
    1. Cloudera manager
    2. HDFS
    3. Flume
    4. Sqoop
    5. Mapreduce
    6. Camus*

Wykład

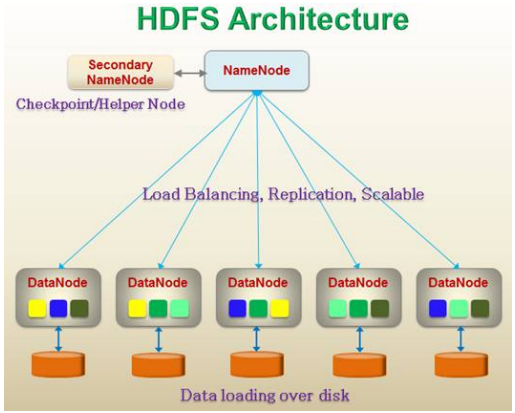
Czym jest hadoop



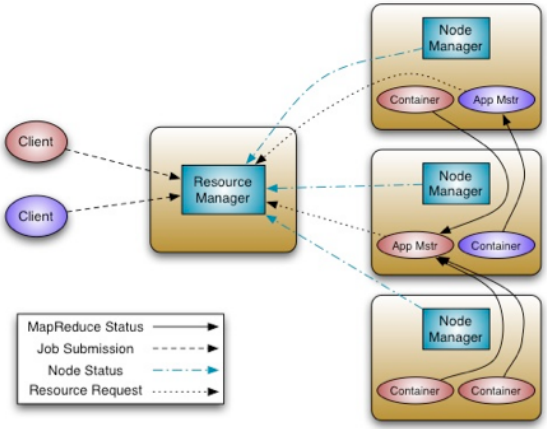
Komponenty


The Apache Hadoop Stack
Hadoop User Experience (HUE)
HUE
Data Exchange
Sqoop
Flume
Log Control
Zoo Keeper
Coordination
Pig
Scripting
Hive
SQL
HIVE
Mahout
ML
mahout
Oozie
Workflow
OOZIE
APACHE HBASE
Hbase
Columnar data store
YARN/Map Reduce V2
Hadoop Distributed File System

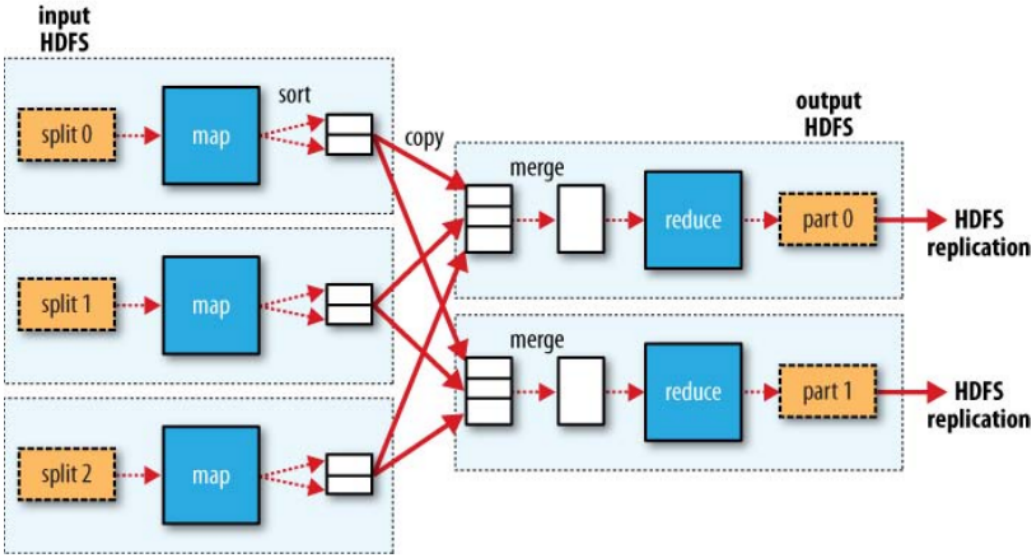
Hdfs


HDFS Architecture
Secondary NameNode
NameNode
Checkpoint/Helper Node
Load Balancing, Replication, Scalable
DataNode
DataNode
DataNode
DataNode
DataNode
Data loading over disk

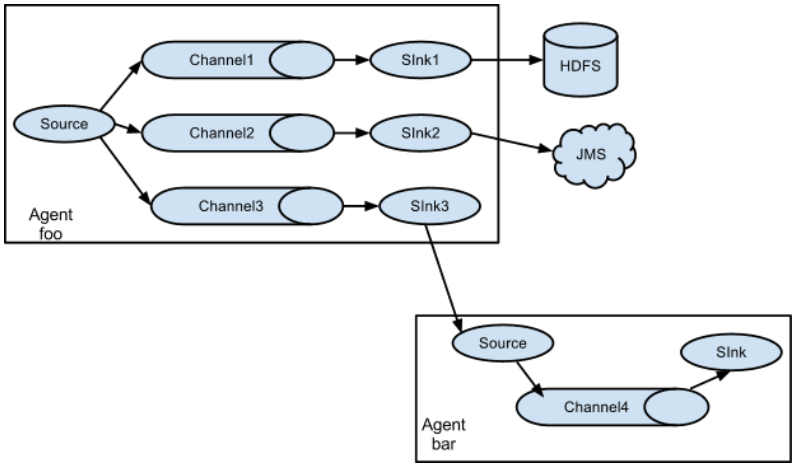
YARN


Client
Client
Resource Manager
Node Manager
Container
App Mstr
Node Manager
App Mstr
Container
Node Manager
Container
Container
MapReduce Status
Job Submission
Node Status
Resource Request

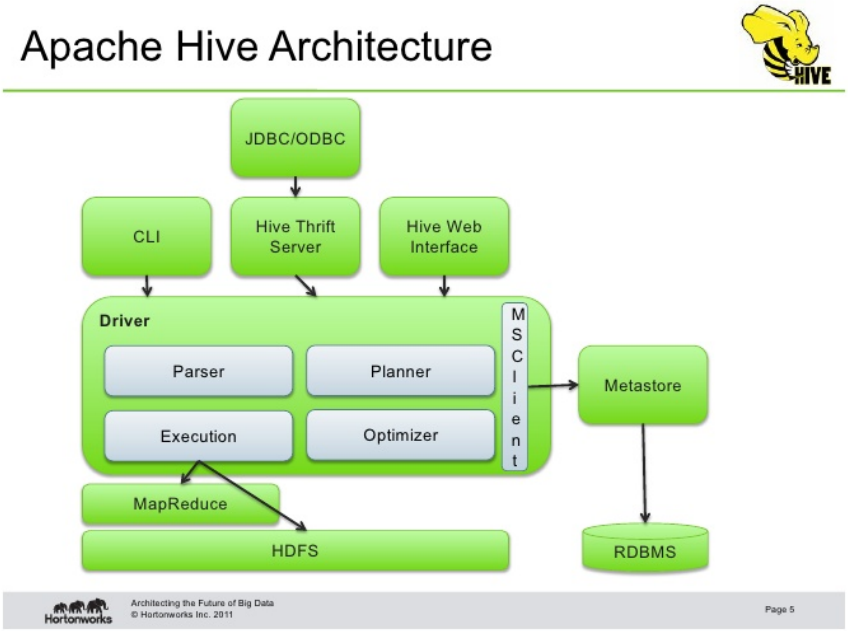
Mapreduce


input
HDFS
split 0
map
sort
split 1
map
split 2
map
copy
merge
reduce
part 0
output
HDFS
HDFS
replication
merge
reduce
part 1
HDFS
replication

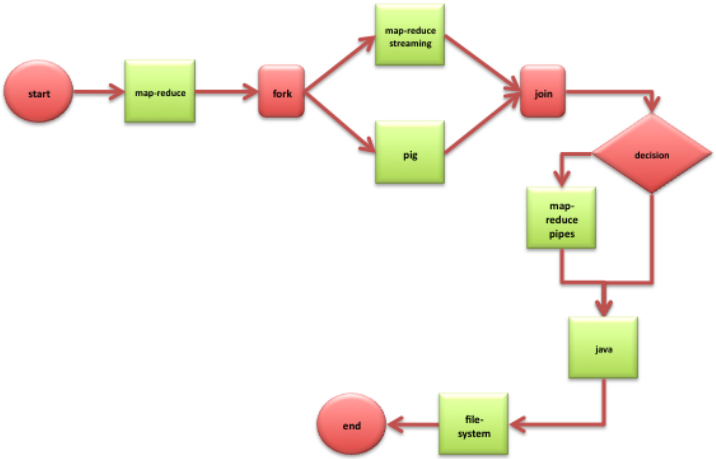
Flume


Channel1
Sink1
HDFS
Source
Channel2
Sink2
JMS
Channel3
Sink3
Agent
foo
Source
Sink
Channel4
Agent
bar

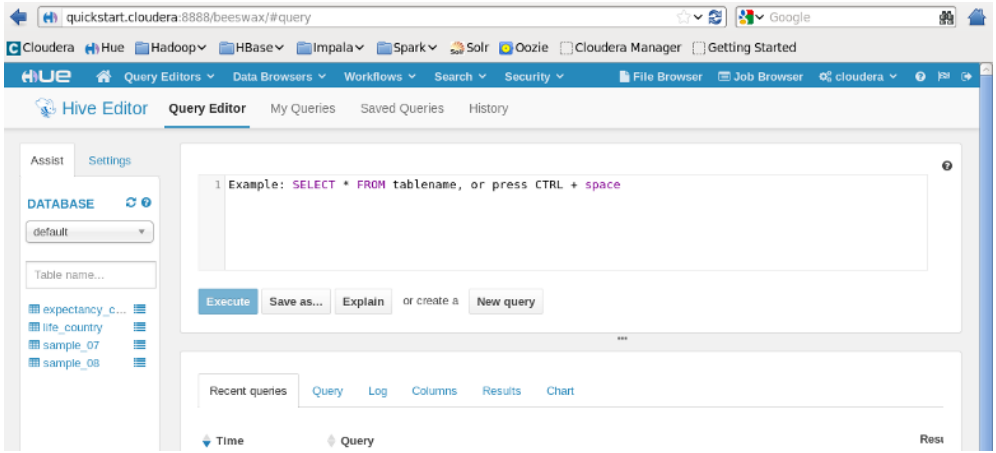
Hive


Apache Hive Architecture
JDBC/ODBC
CLI
Hive Thrift Server
Hive Web Interface
Driver
Parser
Planner
Execution
Optimizer
MS Client
Metastore
MapReduce
HDFS
RDBMS
Architecting the Future of Big Data
© Hortonworks Inc. 2011
Page 5

Oozie


start
map-reduce
fork
map-reduce streaming
pig
join
decision
map-reduce pipes
java
file-system
end

Hue


quickstart.cloudera:8888/beeswax/#query
Google
Cloudera
Hue
Hadoop
HBase
Impala
Spark
Solr
Oozie
Cloudera Manager
Getting Started
Query Editors
Data Browsers
Workflows
Search
Security
File Browser
Job Browser
cloudera
Hive Editor
Query Editor
My Queries
Saved Queries
History
Assist
Settings
DATABASE
default
Table name...
expectancy_c...
life_country
sample_07
sample_08
1 Example: SELECT * FROM tablename, or press CTRL + space
Execute
Save as...
Explain
or create a
New query
Recent queries
Query
Log
Columns
Results
Chart
Time
Query

Dystrybucje

cloudera

MAPR™ TECHNOLOGIES

Hortonworks

Zajęcia praktyczne

- https://github.com/jpodeszwik/Fake-Apache-Log-Generator.git
- https://github.com/jpodeszwik/hadoop-workshop-day1.git

Cloudera manager

HDFS

- vagrant:vagrant /user/vagrant
- vagrant:vagrant /user/vagrant/inputs,outputs
- flume:flume /user/flume
- kursant:kursant /user/kursant
- wrzuć pliki loremipsum i apache_logs do /user/vagrant/inputs

użyj zmienej HADOOP_USER_NAME
hdfs - root

HDFS

- zmień prawa do odczytu do katalogu /user/kursant tak, żeby tylko kursant mógł go odczytać
- spróbuj odczytać katalog jako user 'vagrant'
- spróbuj odczytać katalog jako user 'hdfs'

Snakebite

- utwórz pusty plik w katalogu /user/kursant
- przeczytaj plik /user/vagrant/apache_logs

użyj zmiennej HADOOP_CONF_DIR

webhdfs

- wylistuj katalog
- utwórz plik, zapisz do niego jakieś dane
- usuń przed chwilą utworzony plik

Flume

Mapreduce

Sqoop

- Zainstaluj na vm-cluster-node1 server mysql (sudo apt-get install mysql-server)
- zaimportuj do niego dane z pliku dump.sql (mysql -u root < dump.sql)
- zaimportuj dane z tej bazy na hdfs

Mapreduce

Camus*

Pytania?