Big Data

Il petrolio del presente e del futuro

Fabrizio Soppelsa

Linux day, Belluno, 26 ottobre 2019

Red Hat | IBM

Agenda

- Big data: quanto grande?
- Riflessione non-tecnica
- Riflessione tecnica
  - Architettura Lambda

Big Data: *quanto* grande?

1 byte = 8 bit
Kilo 1024^1 = 1024 bytes
Mega 1024^2 = 1 048 576 bytes
Giga 1024^3 = 1 073 741 824 bytes
Tera 1024^4 = 1 099 511 627 776 bytes
Peta 1024^5 = 1 125 899 906 842 624 bytes
Exa 1024^6 = 1 152 921 504 606 846 976 bytes
Zetta 1024^7 = ecc. ecc.
Yotta 1024^8 = ecc. ecc.

Due conti

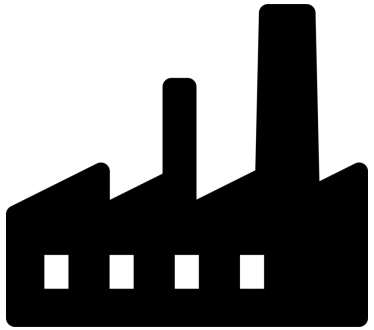
8 bit = A
1 KiloByte: campeggio1998.jpg
1 MegaByte: Cuore di cane
1 GigaByte: From the New World Op. 95
1 TeraByte: Hubble in 1 mese
1 PetaByte: Avatar rendering *totale*
2.5 PetaByte: La mente umana (*stima*)
1 ExaByte: 11 milioni di film in 4K
1 ZettaByte[1]: Tutto il traffico Internet nel 2016

[1] ZettaByte era: https://en.wikipedia.org/wiki/Zettabyte_Era

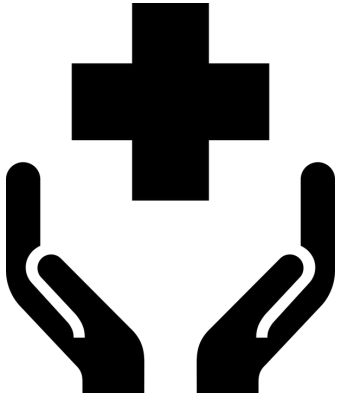
Per capirsi

Chi ha bisogno di Big Data?

Governi
Frodi & evasione fiscale
Ricerca scientifica
Controllo

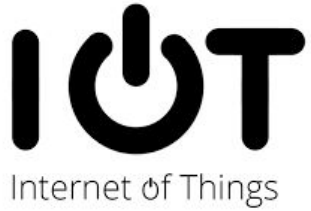


Industria
Qualità prodotto
R&D
Previsione esigenze future



Sanità
Medicina personalizzata
Diagnosi anticipate

Media & Social
Crescita dati utenti
Pubblicità
Tenere alto interesse



IoT
UX personalizzata

Red Hat | IBM

E poi ci siamo noi

Nel 2020

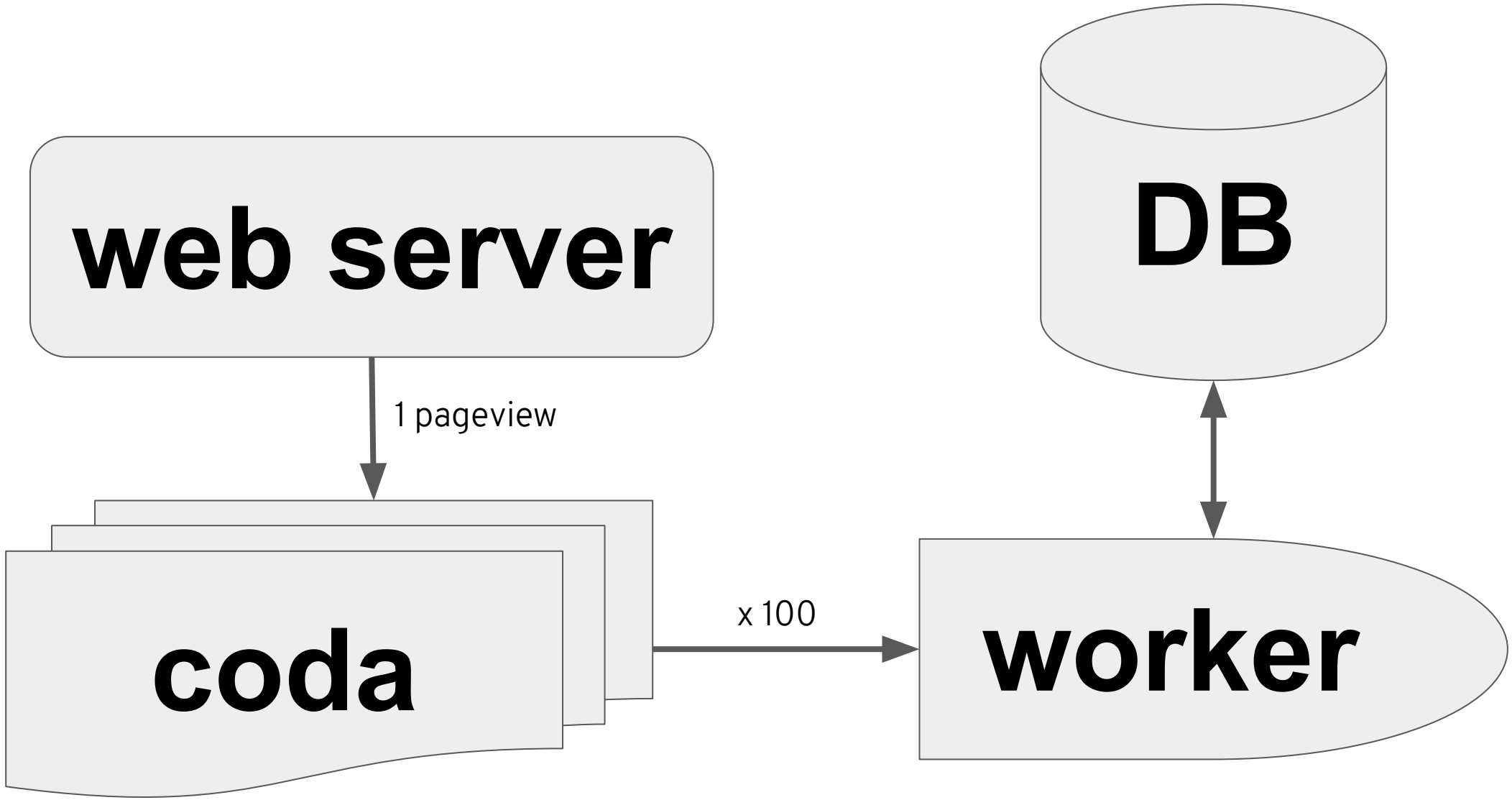
- 6.1 miliardi di utenti di smartphone
- 21 miliardi di dispositivi IoT
- 44 ZettaByte di dati
- 1.7 MegaByte creati da ogni utente ogni secondo

Progettare il Big Data

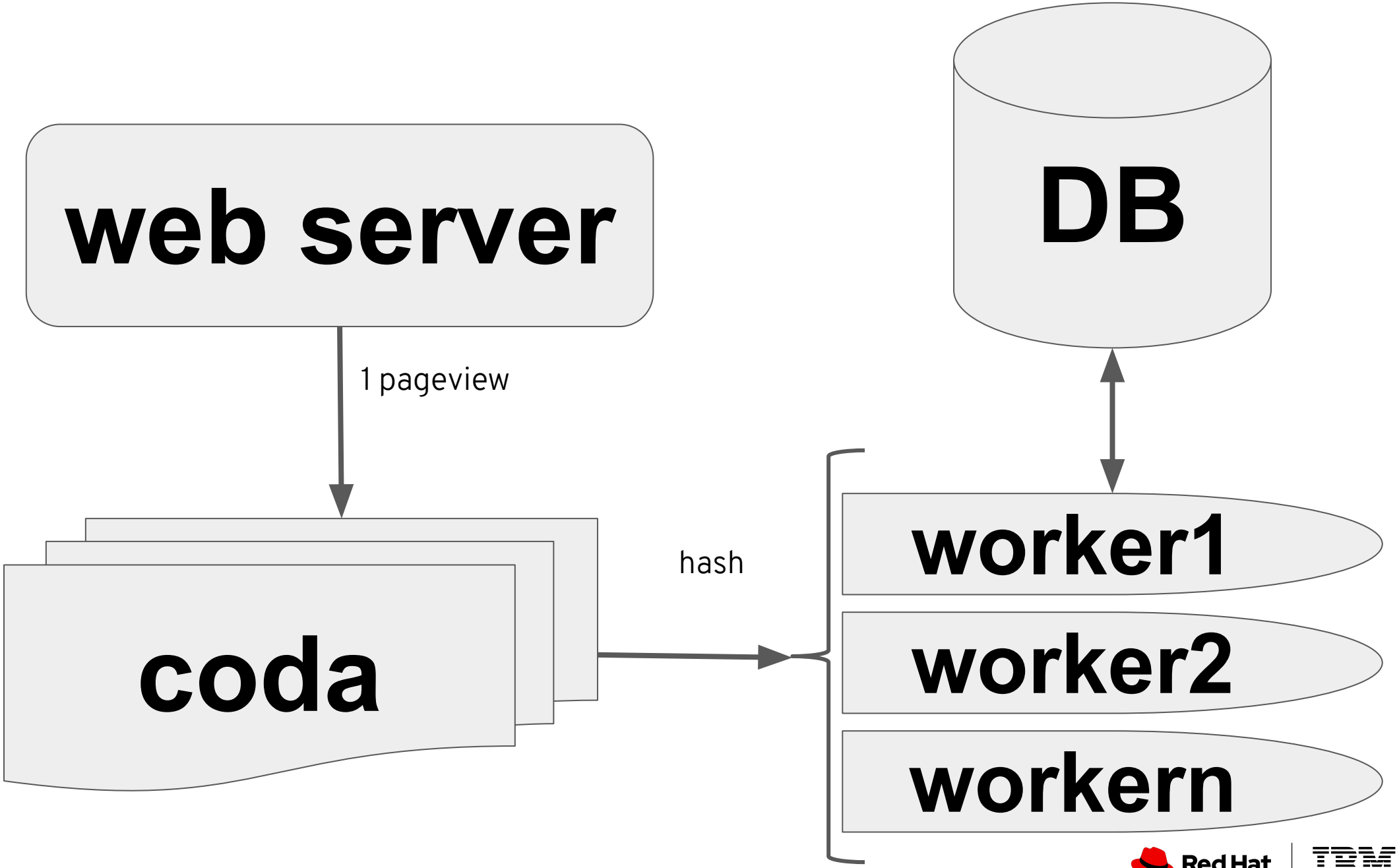
Un tentativo

Colonna	Tipo
id	integer autoincrement
url	varchar(255)
user_id	integer
pageviews	bigint

"Timeout error on inserting to the database"



"Worker crashed"



Red Hat | IBM

```
WORKER 38
Hard Disk Error
Please run the Hard Disk test in System Diagnostic
Hard Disk # (881)

F2 - System Diagnostic
```


for (i = 0; i < 3; i++) {
unsigned int cp_count;
unsigned int len = n;
if (user(group_info[i]){
return -EFAULT;
groupist += NGS_PER_BLOCK;
count -= cp_count;
group_info = kmalloc(usr)
if (!group_info)
return NULL;
group_info->ngroups = gidsetsize;
group_info->nblocks = nblocks;
atomic_set(&group_info->usage, 1)
if (gidsetsize <= NGROUPS_SMALL)
group_info->blocks[0] = gr
else {
for (j = 0; j < nb; j++) {
gid_t *b;
b = (void *)__get(GFP);
if (!b)
goto partial_allocs
group_info->blo[i] = b;
}
struct group_in
int nblocks;
void groups_free(struct=) {
if (group_info->blocks[0]) {
int i;
(i = 0; i < 1; i++)
free_page();
e(group_info);

Red Hat | IBM

Bug: Pageviews * 2

Cosa non ha funzionato?

- Resistenza agli errori?
- Latenza?
- Generalità?
- Prestazioni?
- Incrementabilità?
- Ignoranza sulle proprietà dei dati!

I dati sono grezzi

Spesso non si sa in anticipo cosa si vuole sapere

I dati non cambiano

Non si cancellano

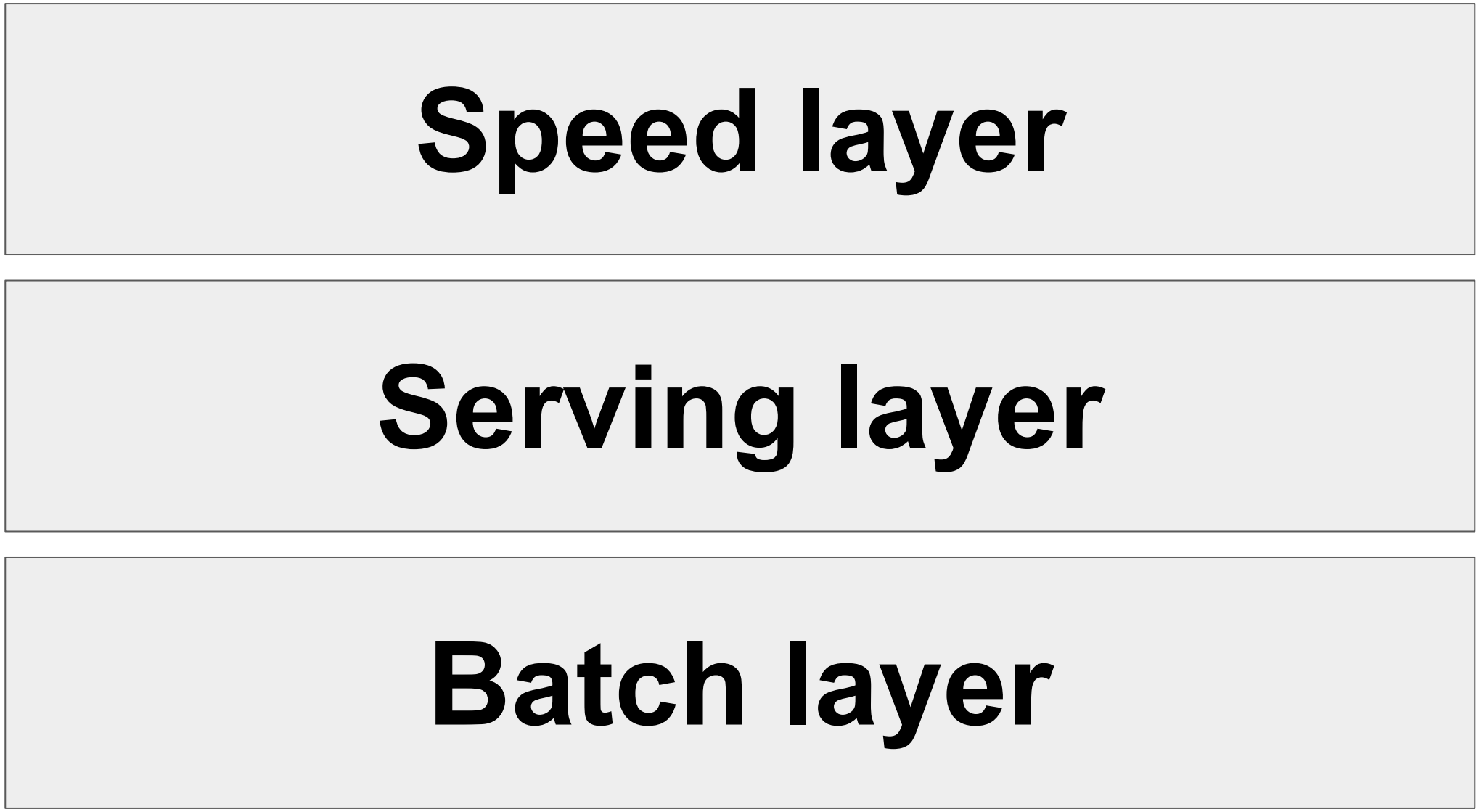
I dati sono eternamente veri

Torino è la capitale d'Italia

Torino è la capitale d'Italia nel 1862

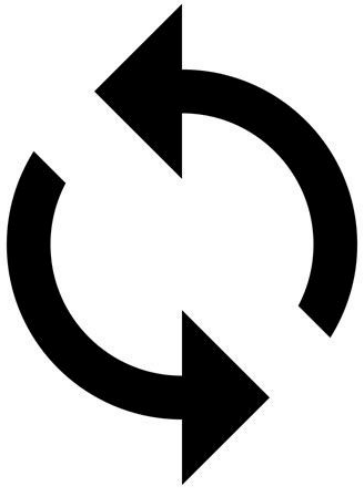
E quindi?

Architettura Lambda



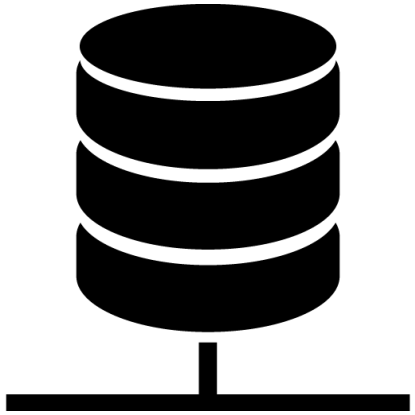
Batch layer

batch view = function(all data)



Serving layer

view = load_views(batch)



Speed layer

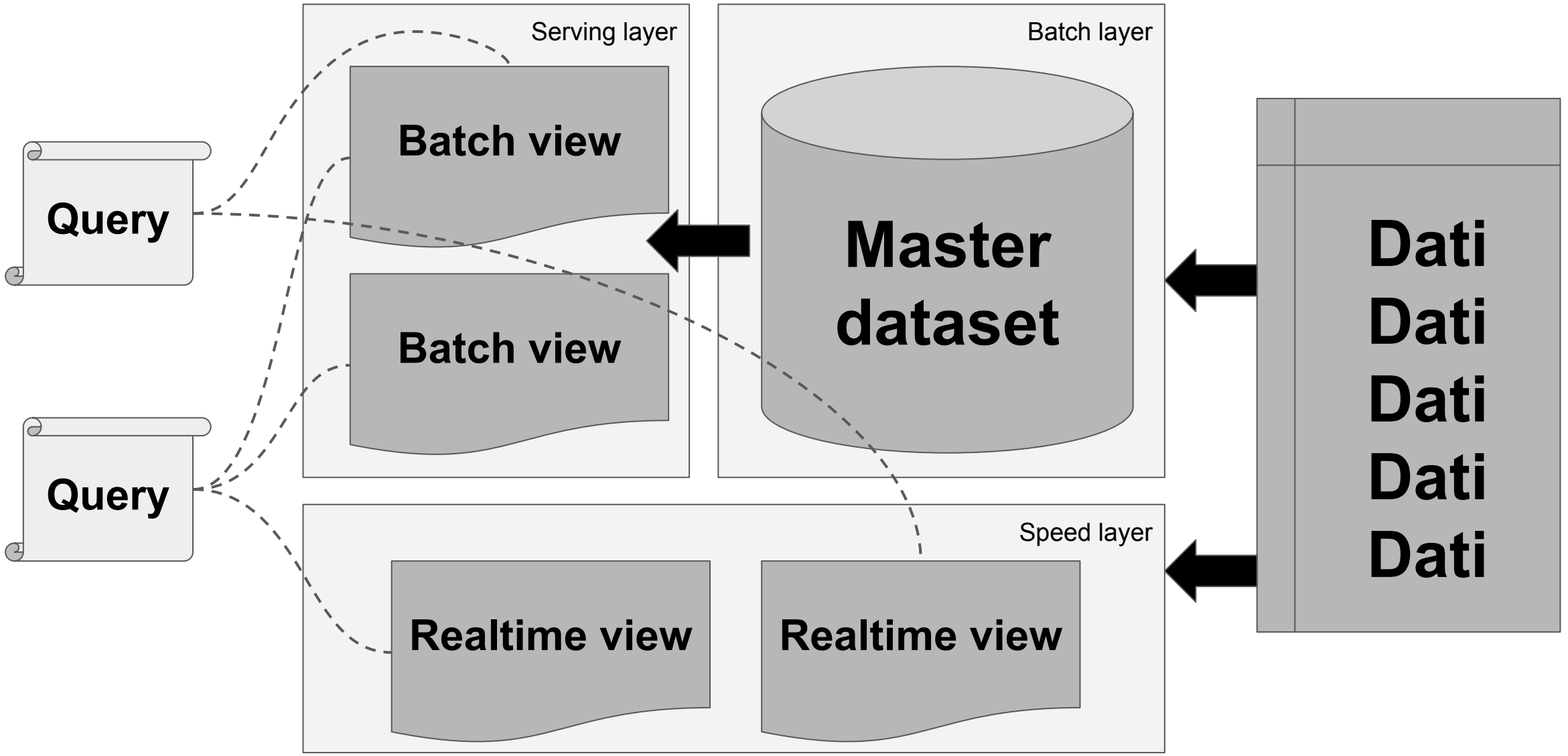
realtime view = function(realtime view, new data)

Quindi

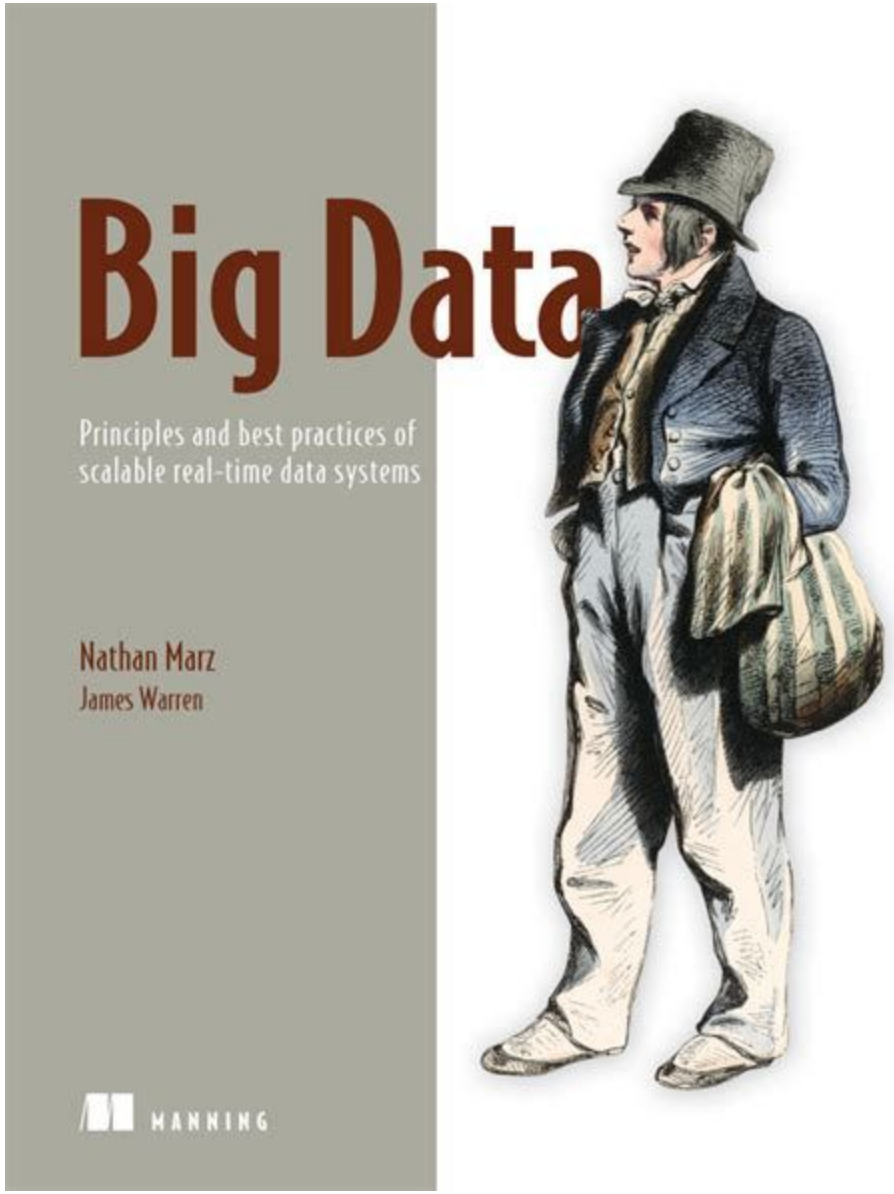
batch view = function(all data)

realtime view = function(realtime view, new data)

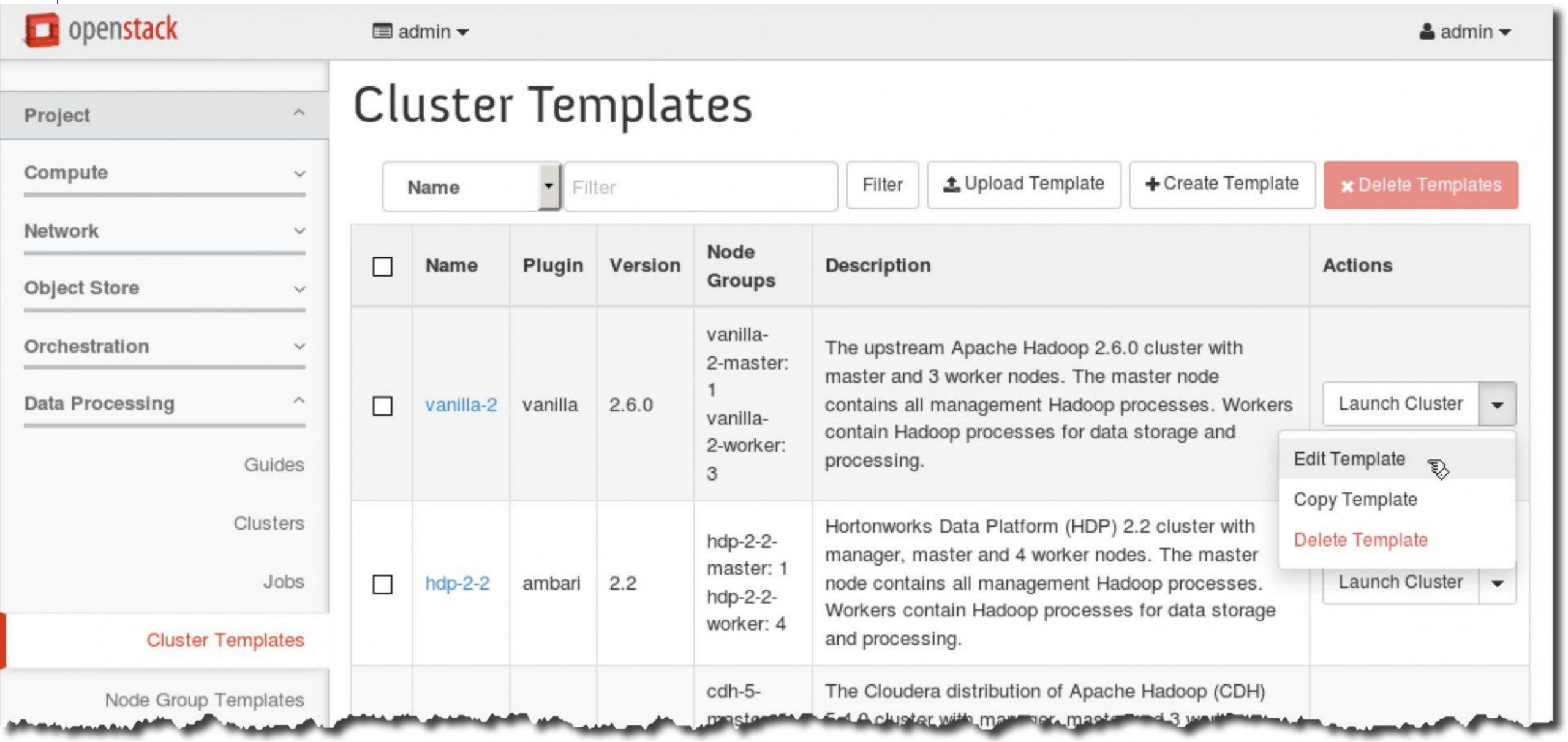
query = function(batch view, realtime view)


Serving layer
Batch layer
Batch view
Batch view
Query
Query
Master dataset
Dati
Dati
Dati
Dati
Dati
Speed layer
Realtime view
Realtime view

Red Hat | IBM


Big Data
Principles and best practices of scalable real-time data systems
Nathan Marz
James Warren
MANNING

Red Hat | IBM


openstack
admin
admin
Project
Compute
Network
Object Store
Orchestration
Data Processing
Guides
Clusters
Jobs
Cluster Templates
Node Group Templates
Cluster Templates
Name
Filter
Filter
Upload Template
Create Template
Delete Templates
Name
Plugin
Version
Node Groups
Description
Actions
vanilla-2
vanilla
2.6.0
vanilla-2-master: 1
vanilla-2-worker: 3
The upstream Apache Hadoop 2.6.0 cluster with master and 3 worker nodes. The master node contains all management Hadoop processes. Workers contain Hadoop processes for data storage and processing.
Launch Cluster
Edit Template
Copy Template
Delete Template
hdp-2-2
ambari
2.2
hdp-2-2-master: 1
hdp-2-2-worker: 4
Hortonworks Data Platform (HDP) 2.2 cluster with manager, master and 4 worker nodes. The master node contains all management Hadoop processes. Workers contain Hadoop processes for data storage and processing.
Launch Cluster
cdh-5-
The Cloudera distribution of Apache Hadoop (CDH)

Red Hat | IBM

Grazie

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.

linkedin.com/company/red-hat

youtube.com/user/RedHatVideos

facebook.com/redhatinc

twitter.com/RedHat

Red Hat | IBM