R을 활용한 Big Data 분석

NexR 데이터 분석팀
권정민
irene.kwon@nexr.com

Big Data

Volume
Variety
Velocity

IBM Homepage (2011)

If I'd known they wanted me to use all this info – I would never have asked for it!
Julie

기존 방법:

R, SAS, SPSS, Excel,...

SingleCore Processing

In-Memory

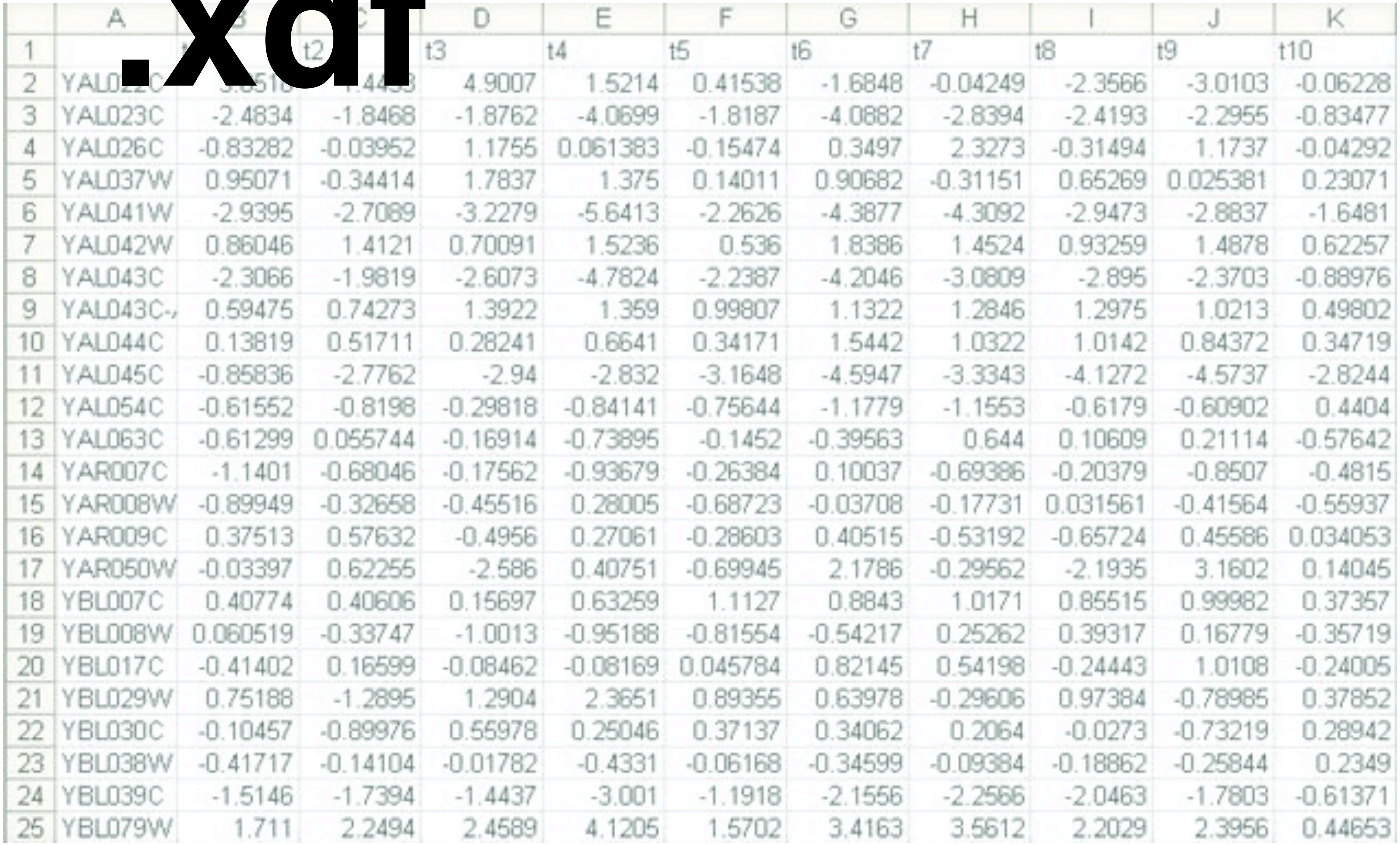
Sampling, Aggregation

R with Modification

RevolutionR

RevoScaleR

.xdf



	A	B	C	D	E	F	G	H	I	J	K
1		t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
2	YAL022C	3.0516	1.4453	4.9007	1.5214	0.41538	-1.6848	-0.04249	-2.3566	-3.0103	-0.06228
3	YAL023C	-2.4834	-1.8468	-1.8762	-4.0699	-1.8187	-4.0882	-2.8394	-2.4193	-2.2955	-0.83477
4	YAL026C	-0.83282	-0.03952	1.1755	0.061383	-0.15474	0.3497	2.3273	-0.31494	1.1737	-0.04292
5	YAL037W	0.95071	-0.34414	1.7837	1.375	0.14011	0.90682	-0.31151	0.65269	0.025381	0.23071
6	YAL041W	-2.9395	-2.7089	-3.2279	-5.6413	-2.2626	-4.3877	-4.3092	-2.9473	-2.8837	-1.6481
7	YAL042W	0.86046	1.4121	0.70091	1.5236	0.536	1.8386	1.4524	0.93259	1.4878	0.62257
8	YAL043C	-2.3066	-1.9819	-2.6073	-4.7824	-2.2387	-4.2046	-3.0809	-2.895	-2.3703	-0.88976
9	YAL043C-/	0.59475	0.74273	1.3922	1.359	0.99807	1.1322	1.2846	1.2975	1.0213	0.49802
10	YAL044C	0.13819	0.51711	0.28241	0.6641	0.34171	1.5442	1.0322	1.0142	0.84372	0.34719
11	YAL045C	-0.85836	-2.7762	-2.94	-2.832	-3.1648	-4.5947	-3.3343	-4.1272	-4.5737	-2.8244
12	YAL054C	-0.61552	-0.8198	-0.29818	-0.84141	-0.75644	-1.1779	-1.1553	-0.6179	-0.60902	0.4404
13	YAL063C	-0.61299	0.055744	-0.16914	-0.73895	-0.1452	-0.39563	0.644	0.10609	0.21114	-0.57642
14	YAR007C	-1.1401	-0.68046	-0.17562	-0.93679	-0.26384	0.10037	-0.69386	-0.20379	-0.8507	-0.4815
15	YAR008W	-0.89949	-0.32658	-0.45516	0.28005	-0.68723	-0.03708	-0.17731	0.031561	-0.41564	-0.55937
16	YAR009C	0.37513	0.57632	-0.4956	0.27061	-0.28603	0.40515	-0.53192	-0.65724	0.45586	0.034053
17	YAR050W	-0.03397	0.62255	-2.586	0.40751	-0.69945	2.1786	-0.29562	-2.1935	3.1602	0.14045
18	YBL007C	0.40774	0.40606	0.15697	0.63259	1.1127	0.8843	1.0171	0.85515	0.99982	0.37357
19	YBL008W	0.060519	-0.33747	-1.0013	-0.95188	-0.81554	-0.54217	0.25262	0.39317	0.16779	-0.35719
20	YBL017C	-0.41402	0.16599	-0.08462	-0.08169	0.045784	0.82145	0.54198	-0.24443	1.0108	-0.24005
21	YBL029W	0.75188	-1.2895	1.2904	2.3651	0.89355	0.63978	-0.29606	0.97384	-0.78985	0.37852
22	YBL030C	-0.10457	-0.89976	0.55978	0.25046	0.37137	0.34062	0.2064	-0.0273	-0.73219	0.28942
23	YBL038W	-0.41717	-0.14104	-0.01782	-0.4331	-0.06168	-0.34599	-0.09384	-0.18862	-0.25844	0.2349
24	YBL039C	-1.5146	-1.7394	-1.4437	-3.001	-1.1918	-2.1556	-2.2566	-2.0463	-1.7803	-0.61371
25	YBL079W	1.711	2.2494	2.4589	4.1205	1.5702	3.4163	3.5612	2.2029	2.3956	0.44653

In-HDD:

데이터 크기에 구애받지 않음

K-Means

Regression

Aggregation

외부 패키지 함수 사용 불가

내장 함수 부족

낮은 문법 호환성

Bigmemory

bigmemory
biganalytics
bigalgebra
bigtabulate

C++ Pointer:

big.matrix

shared.big.matrix

filebacked.big.matrix

접근 용이
다양한 함수

NexR
TOWARD OPEN PLATFORM

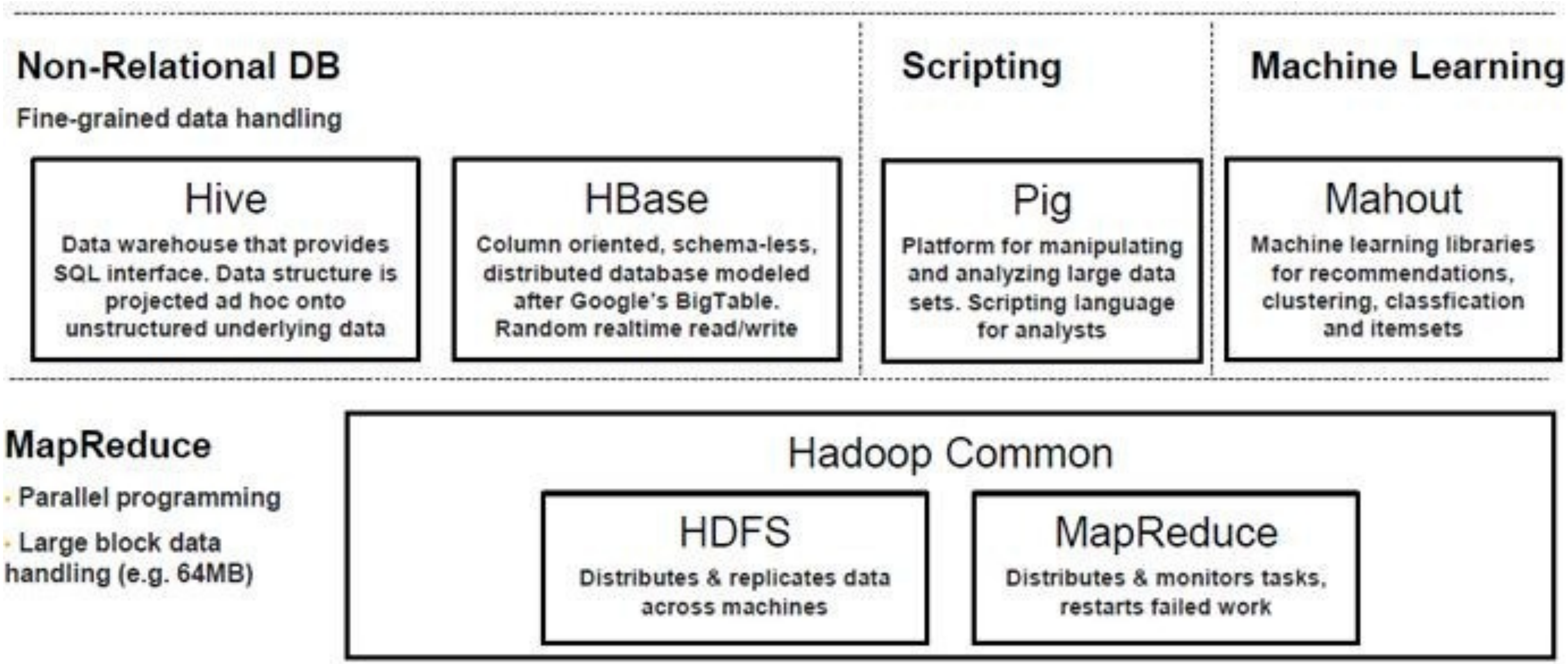
One Object, One Type

Oracle R Enterprise

In-Database:

SQL을 사용하지 않고
DB 내 데이터를 분석

http://hadoop.apache.org/



Non-Relational DB

Fine-grained data handling

Hive

Data warehouse that provides SQL interface. Data structure is projected ad hoc onto unstructured underlying data

HBase

Column oriented, schema-less, distributed database modeled after Google's BigTable. Random realtime read/write

Scripting

Pig

Platform for manipulating and analyzing large data sets. Scripting language for analysts

Machine Learning

Mahout

Machine learning libraries for recommendations, clustering, classfication and itemsets

MapReduce

- Parallel programming
- Large block data handling (e.g. 64MB)

Hadoop Common

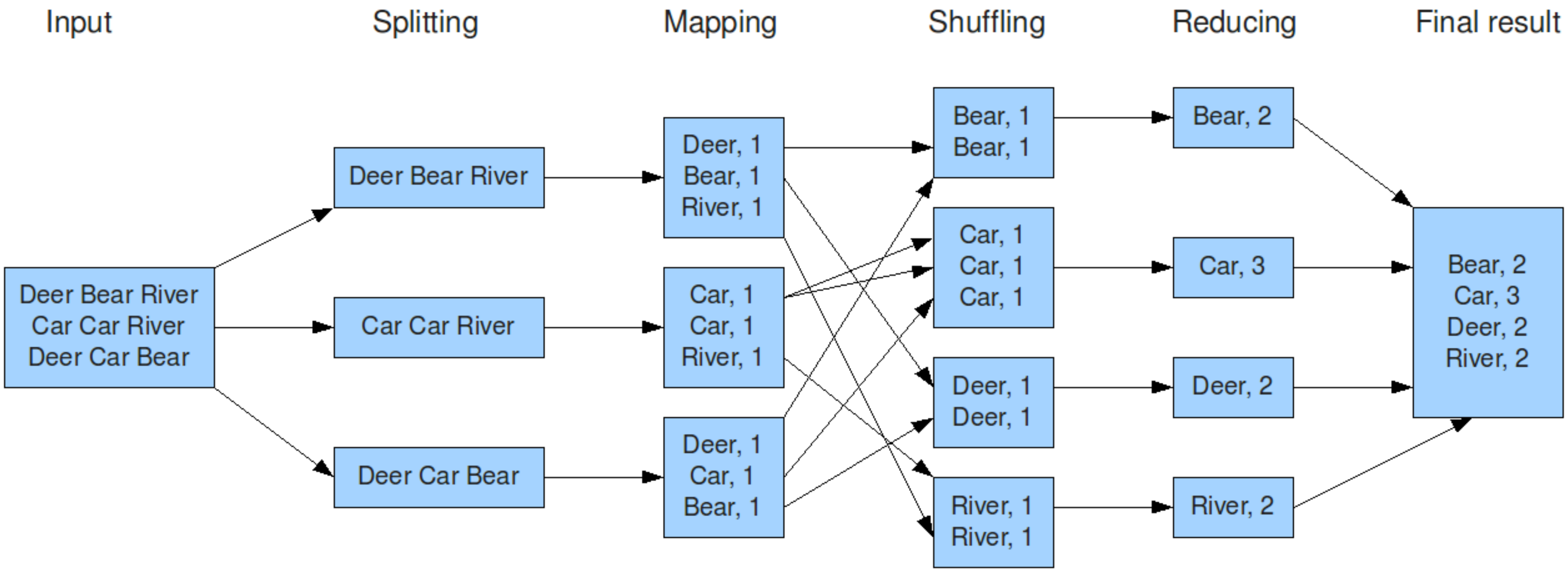
HDFS

Distributes & replicates data across machines

MapReduce

Distributes & monitors tasks, restarts failed work

The overall MapReduce word count process



Input

Deer Bear River
Car Car River
Deer Car Bear

Splitting

Deer Bear River

Car Car River

Deer Car Bear

Mapping

Deer, 1
Bear, 1
River, 1

Car, 1
Car, 1
River, 1

Deer, 1
Car, 1
Bear, 1

Shuffling

Bear, 1
Bear, 1

Car, 1
Car, 1
Car, 1

Deer, 1
Deer, 1

River, 1
River, 1

Reducing

Bear, 2

Car, 3

Deer, 2

River, 2

Final result

Bear, 2
Car, 3
Deer, 2
River, 2

http://www.searchworkings.org/blog/-/blogs/introduction-to-hadoop

IBM Ricardo

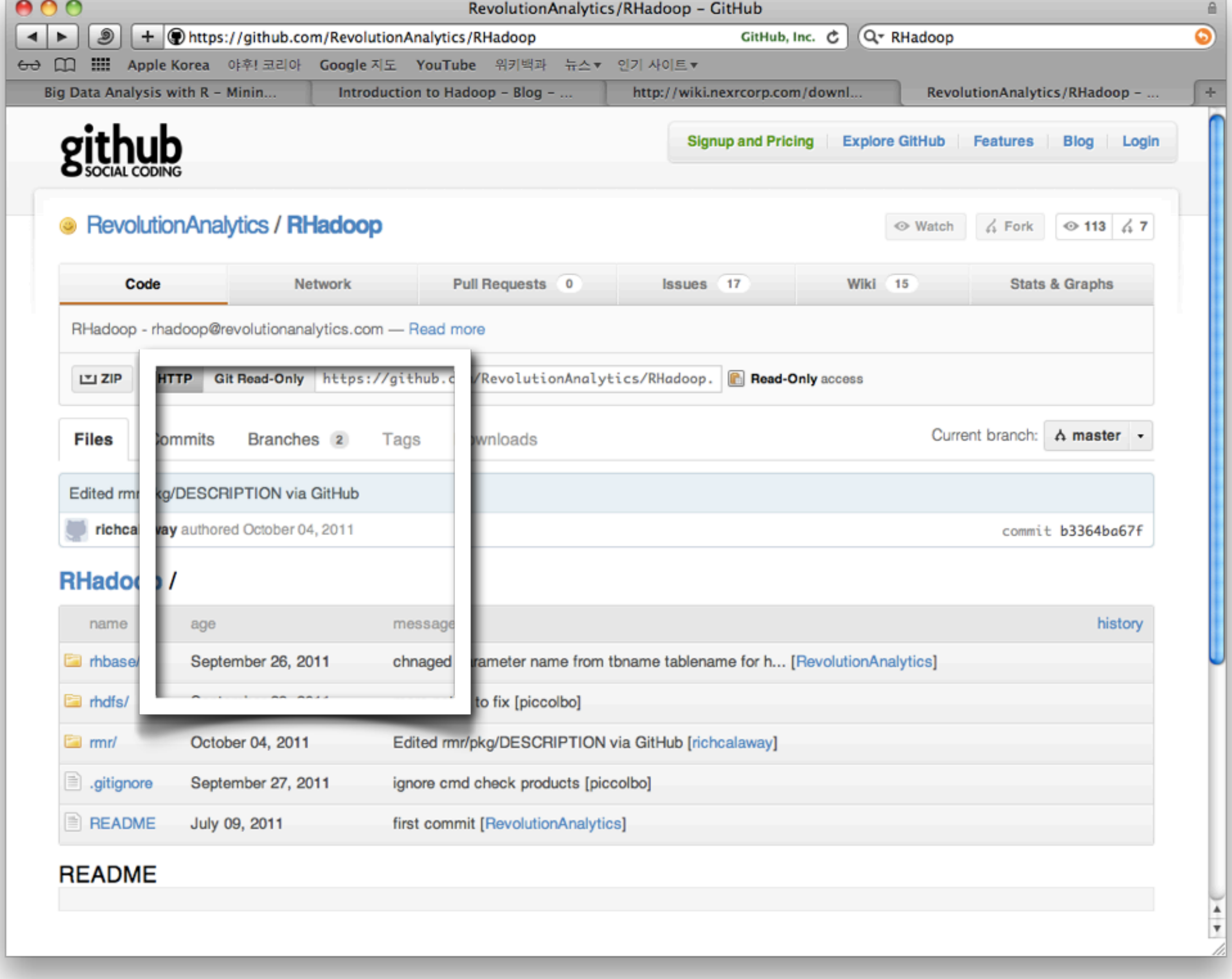
Large Part:

Jaql + Hadoop

Small Part:

R

RHadoop



RevolutionAnalytics / RHadoop

RHadoop - rhadoop@revolutionanalytics.com — Read more

https://github.c /RevolutionAnalytics/RHadoop. Read-Only access

Edited rmr kg/DESCRIPTION via GitHub

richca ay authored October 04, 2011 commit b3364ba67f

RHado /

name	age	message
rhbase/	September 26, 2011	chnaged arameter name from tbname tablename for h... [RevolutionAnalytics]
rhdfs/	[illegible]	to fix [piccolbo]
rmr/	October 04, 2011	Edited rmr/pkg/DESCRIPTION via GitHub [richcalaway]
.gitignore	September 27, 2011	ignore cmd check products [piccolbo]
README	July 09, 2011	first commit [RevolutionAnalytics]

README

Rhipe

(R and Hadoop Integrated Processing Environment)

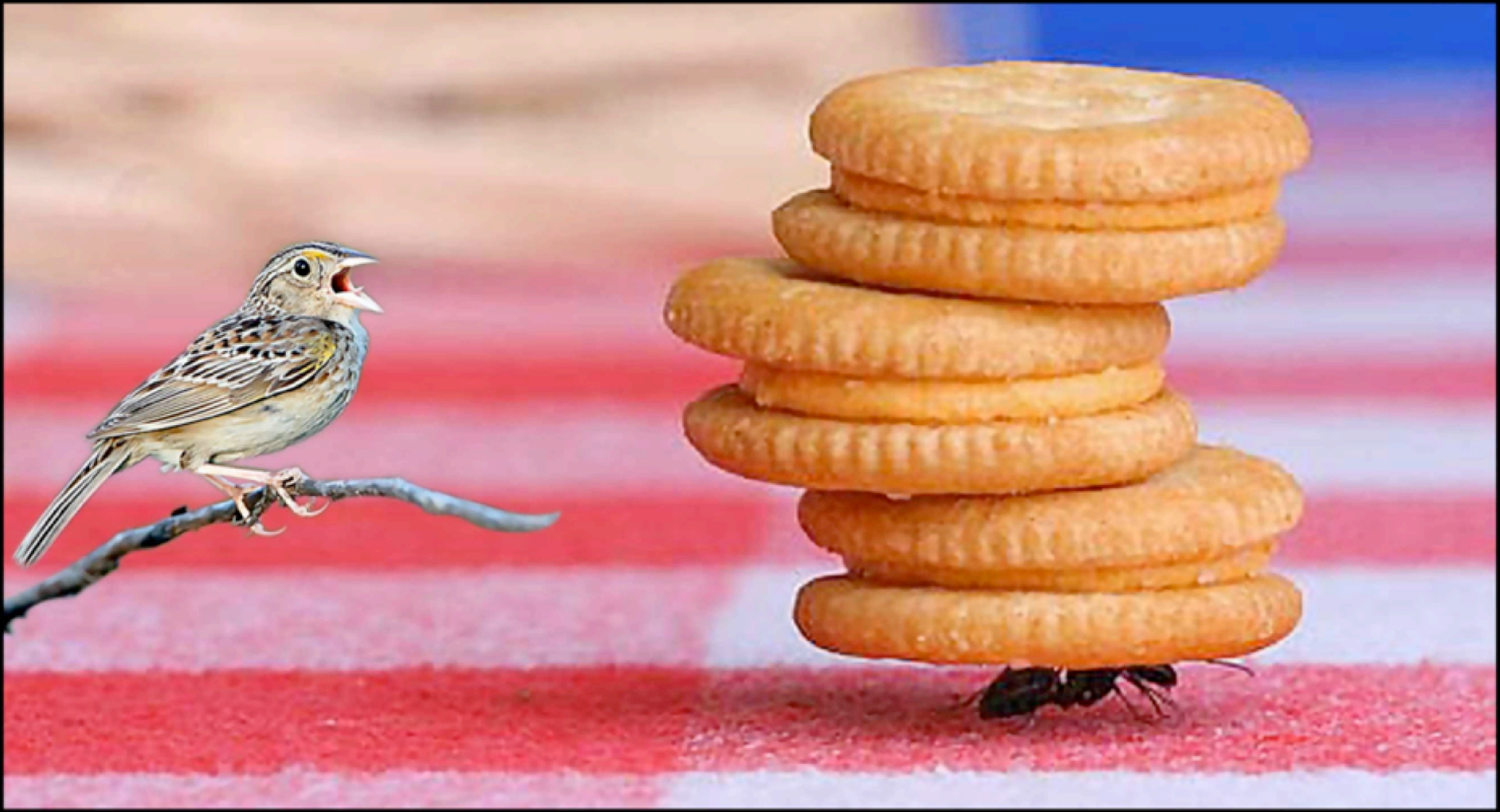
In-Hadoop:

데이터 크기에 구애받지 않음

NEXR
TOWARD OPEN PLATFORM

MapReduce In R:

R에서 바로 Hadoop 데이터 처리



HARDWORK:

it's not easy to get hungry...

RHive

R

RJava

RServe

Hive

Out-Of-Memory

RevolutionR
bigmemory
Oracle R

Ricardo
Rhipe
RHadoop
RHive

Big Data Analysis in R

Q & A

“What is the meaning of life”

42.