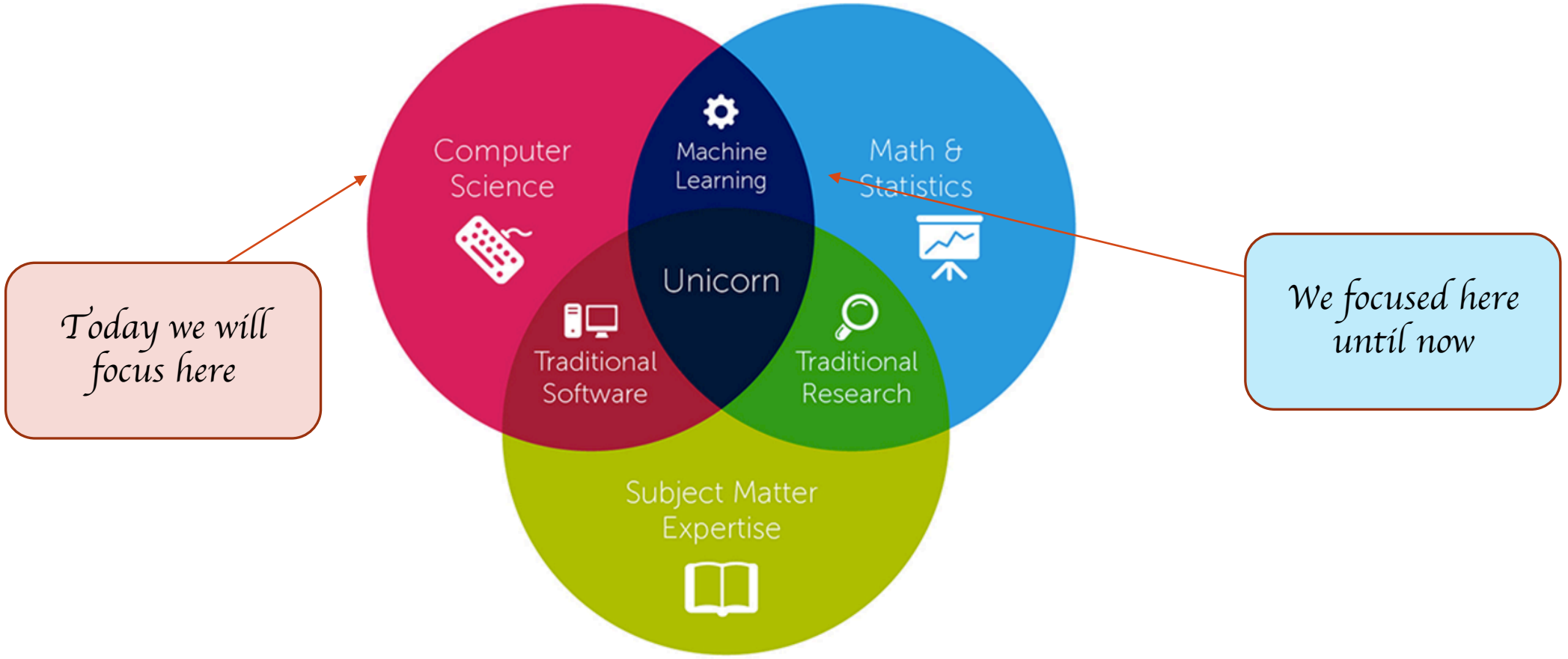
DSLA
DATA SCIENCE LEARNING ACADEMY

Introduction to Big Data

DSLA COURSE

ROHIT PADEBETTU

Who is a Data Scientist?



Time Gained in Flight

Select Airports

Flight origin

JFK

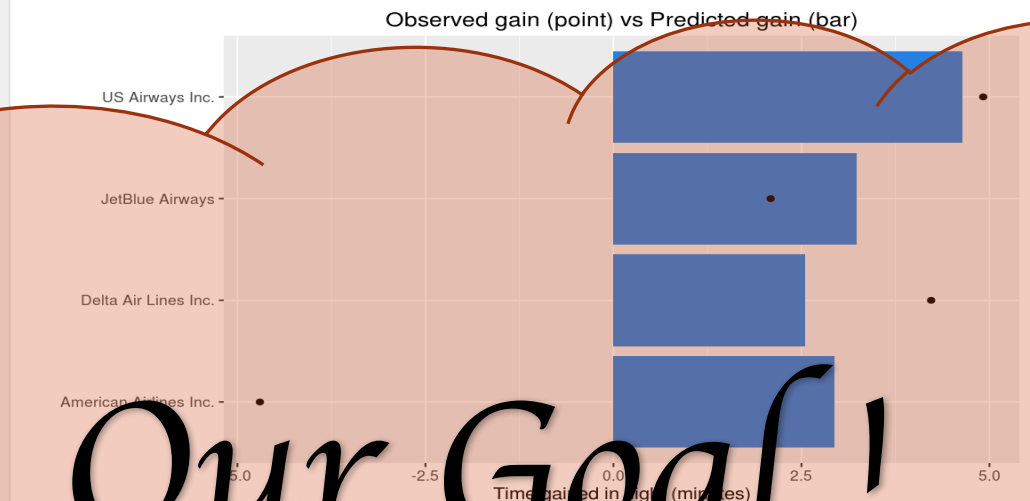
Flight destination

LAS

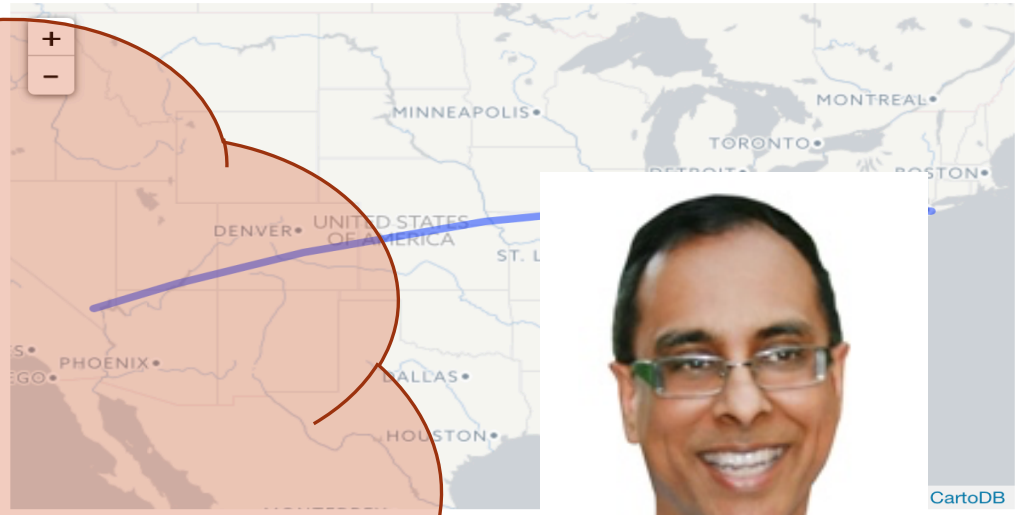
Background

Given that your flight was delayed by 15 minutes or more, what is the likelihood your airline carrier will make up time in route? Some of the most signficant factors for making up time are flight distance and airline carrier. The data model behind this dashboard is based on flights from NYC airports in 2013.

Observed versus predicted time gain



Route



Data details

Show 10 entries

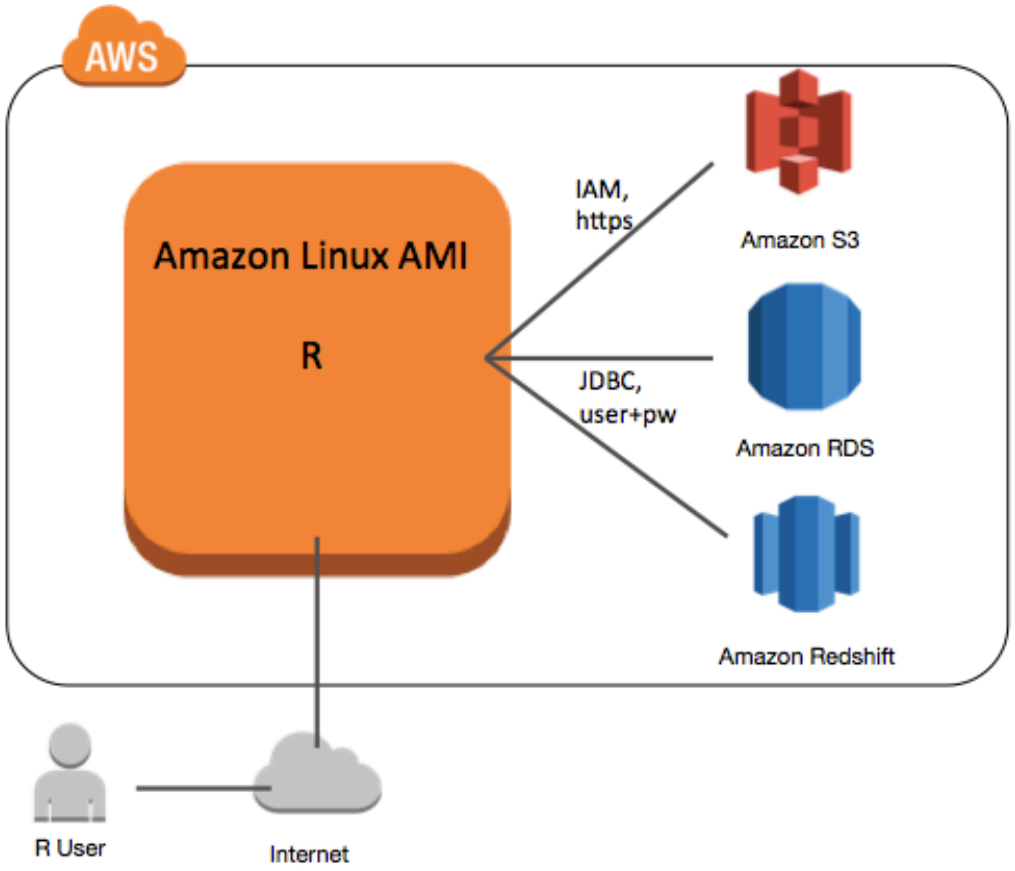
	airline	flights	distance	avg_dep_delay	avg_arr_delay		gain
1	JetBlue Airways	286	2248	61.7	59.6		3.2
2	American Airlines Inc.	97	2248	57.2	61.9	-4.7	2.9
3	Delta Air Lines Inc.	227	2248	58.7	54.4	4.2	2.5
4	US Airways Inc.	192	2248	57.4	52.5	4.9	4.6

Showing 1 to 4 of 4 entries

Previous 1 Next

Our Goal !

AWS Cloud Architecture

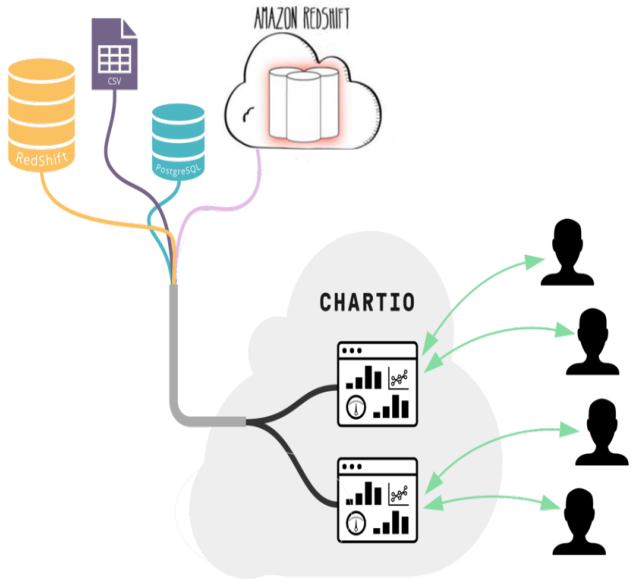


Yotta (Y)	10^{24}	1 septillion
Zetta (Z)	10^{21}	1 sextillion
Exa (E)	10^{18}	1 quintillion
Peta (P)	10^{15}	1 quadrillion
Tera (T)	10^{12}	1 trillion
Giga (G)	10^{9}	1 billion
Mega (M)	10^{6}	1 million
kilo (k)	10^{3}	1 thousand

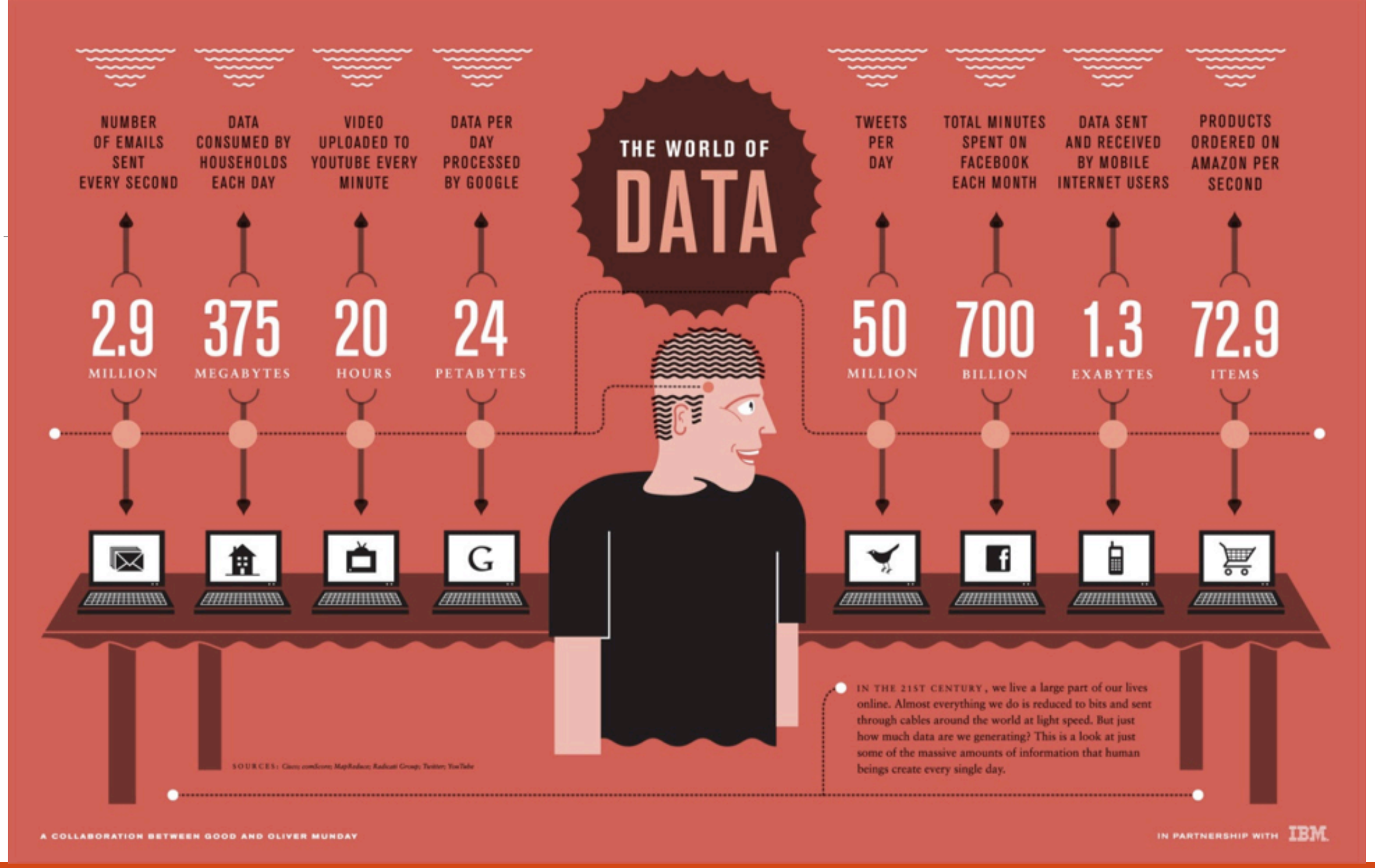
Most modern databases can manage approximately 6TB of data

Data in all Shapes & Sizes

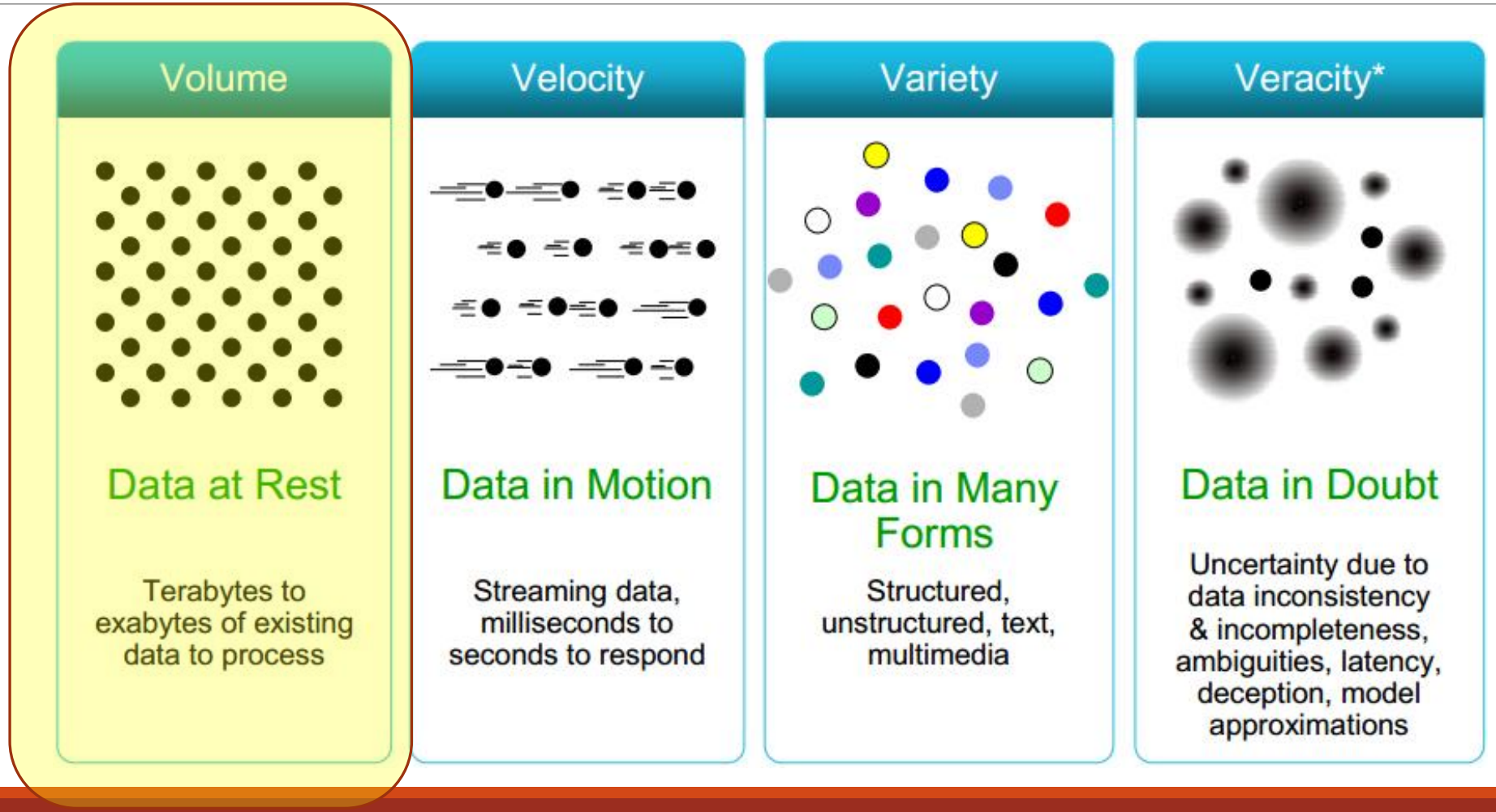
A Data Warehouse is a "Relational Database" that is used primarily for query and analysis rather than transaction processing. (OLAP vs OLTP)



Amazon Redshift is branded as a Petabyte scale database


THE WORLD OF
DATA
NUMBER OF EMAILS SENT EVERY SECOND
2.9
MILLION
DATA CONSUMED BY HOUSEHOLDS EACH DAY
375
MEGABYTES
VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20
HOURS
DATA PER DAY PROCESSED BY GOOGLE
24
PETABYTES
TWEETS PER DAY
50
MILLION
TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700
BILLION
DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3
EXABYTES
PRODUCTS ORDERED ON AMAZON PER SECOND
72.9
ITEMS
IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.
SOURCES:
A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY
IN PARTNERSHIP WITH IBM

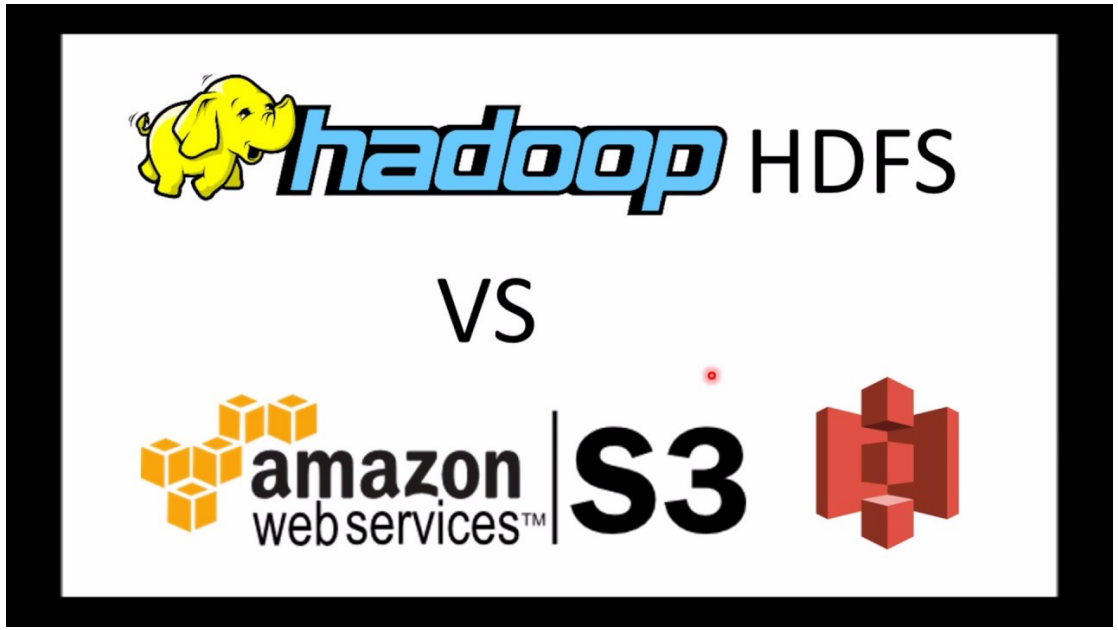
Characteristics of Data



Big Data Will Scale To Exabytes

Store Big Data

How ?



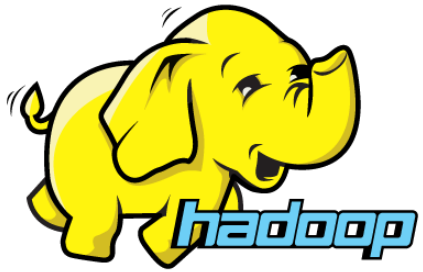
RACKSPACE CLOUD BLOCK STORAGE

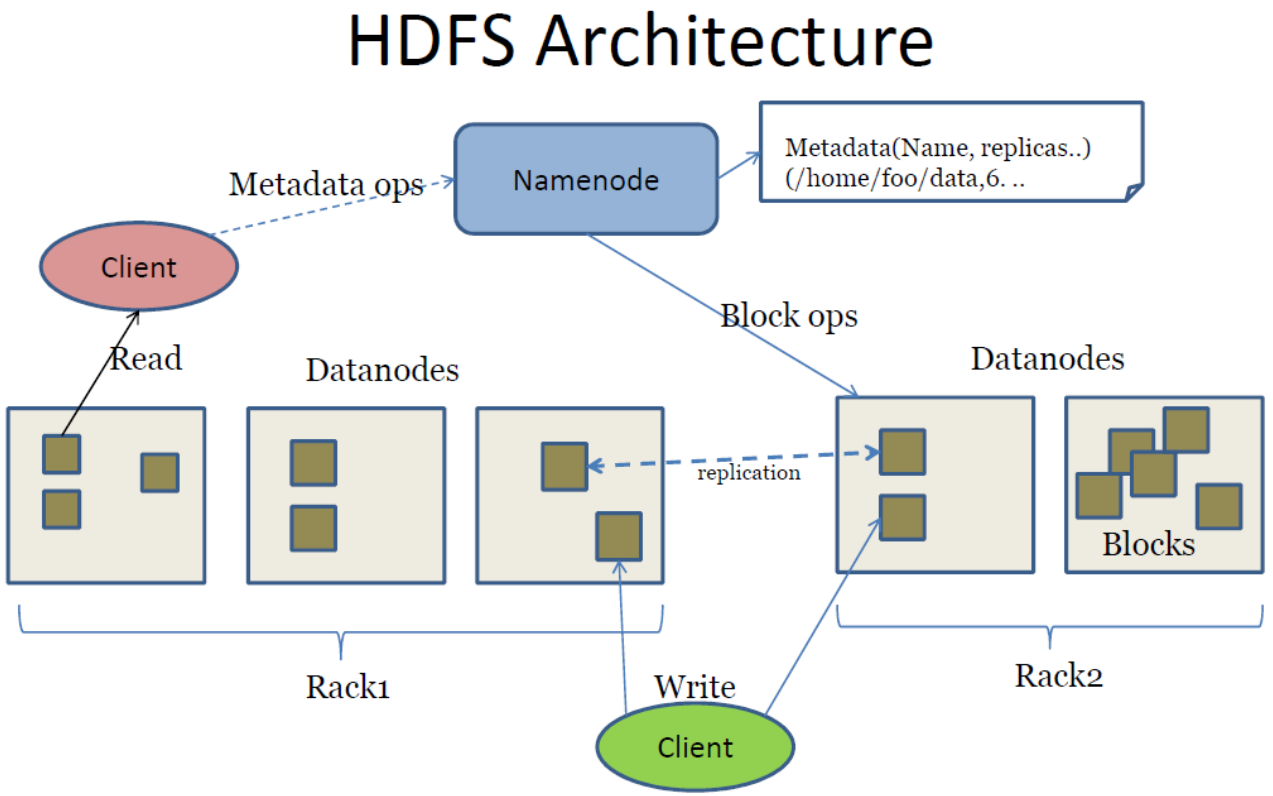
Google Cloud Storage

Microsoft Azure Blob Storage

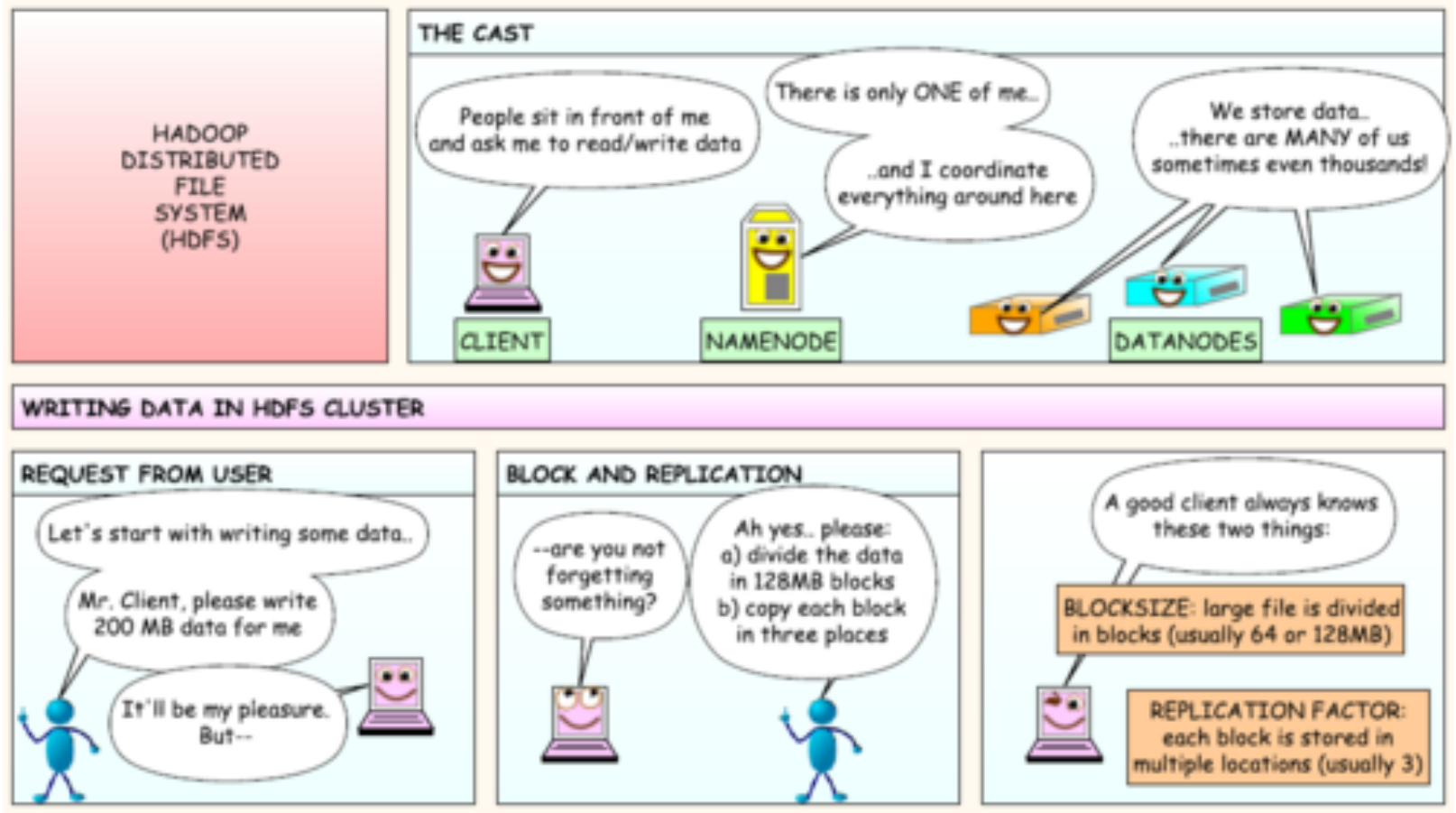
IBM BlueMix

Distributed Storage

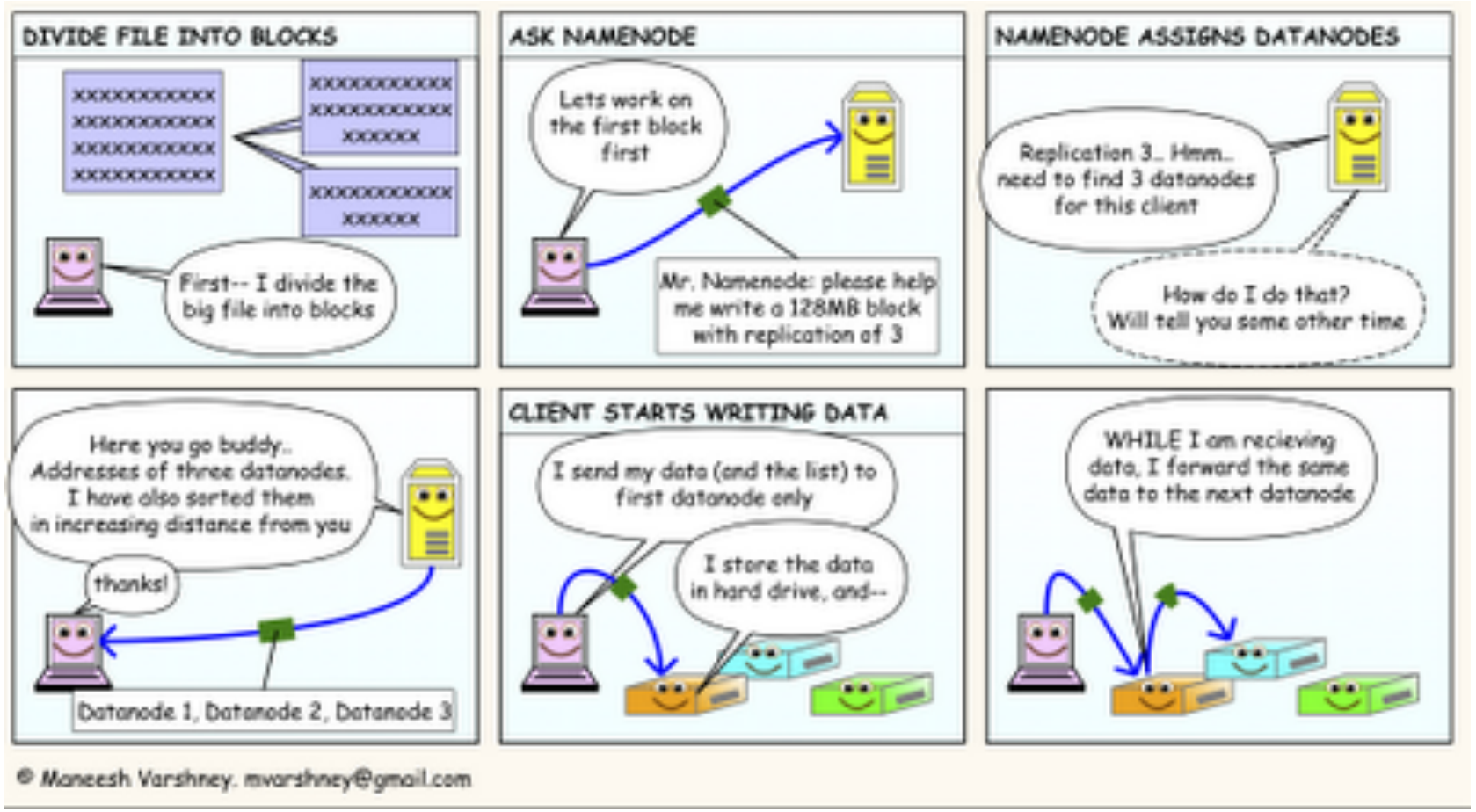




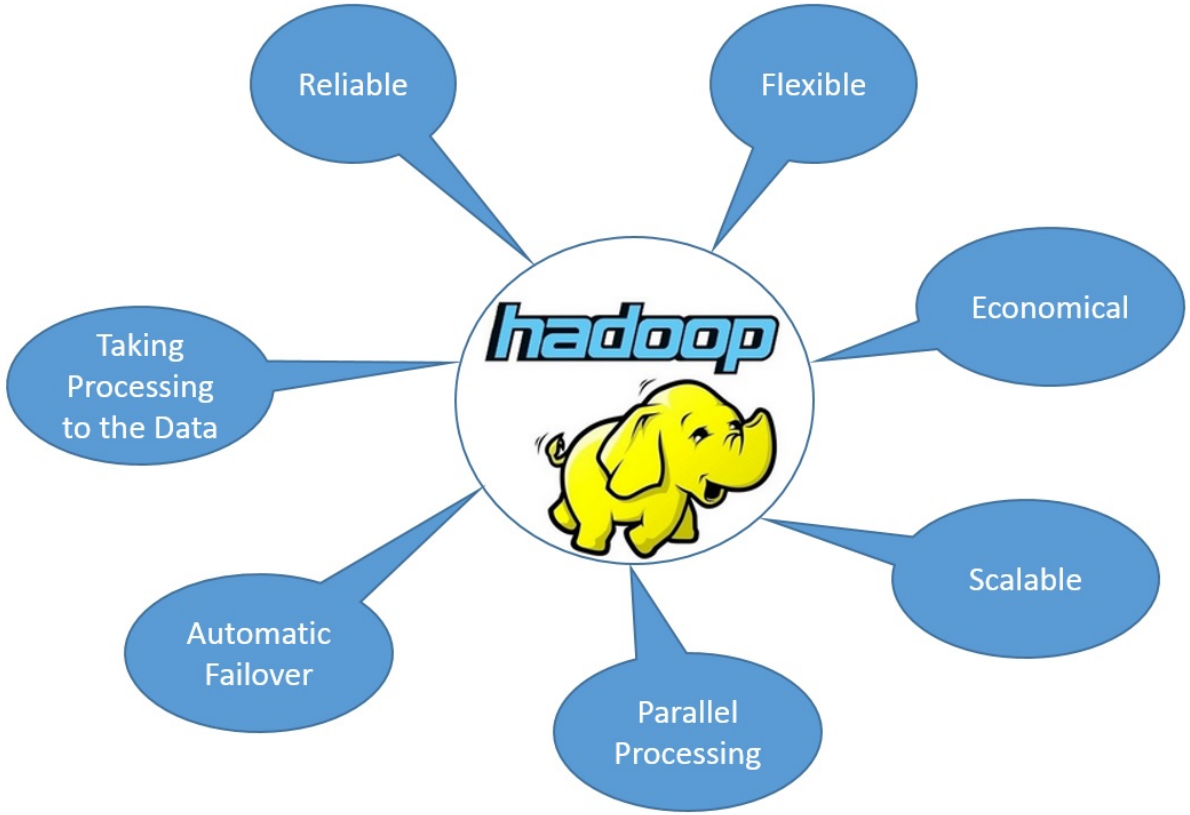
Distributed Storage



Distributed Storage



Hadoop Advantages



Access Distributed Data

DSLA
DATA SCIENCE LEARNING ACADEMY

How?

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

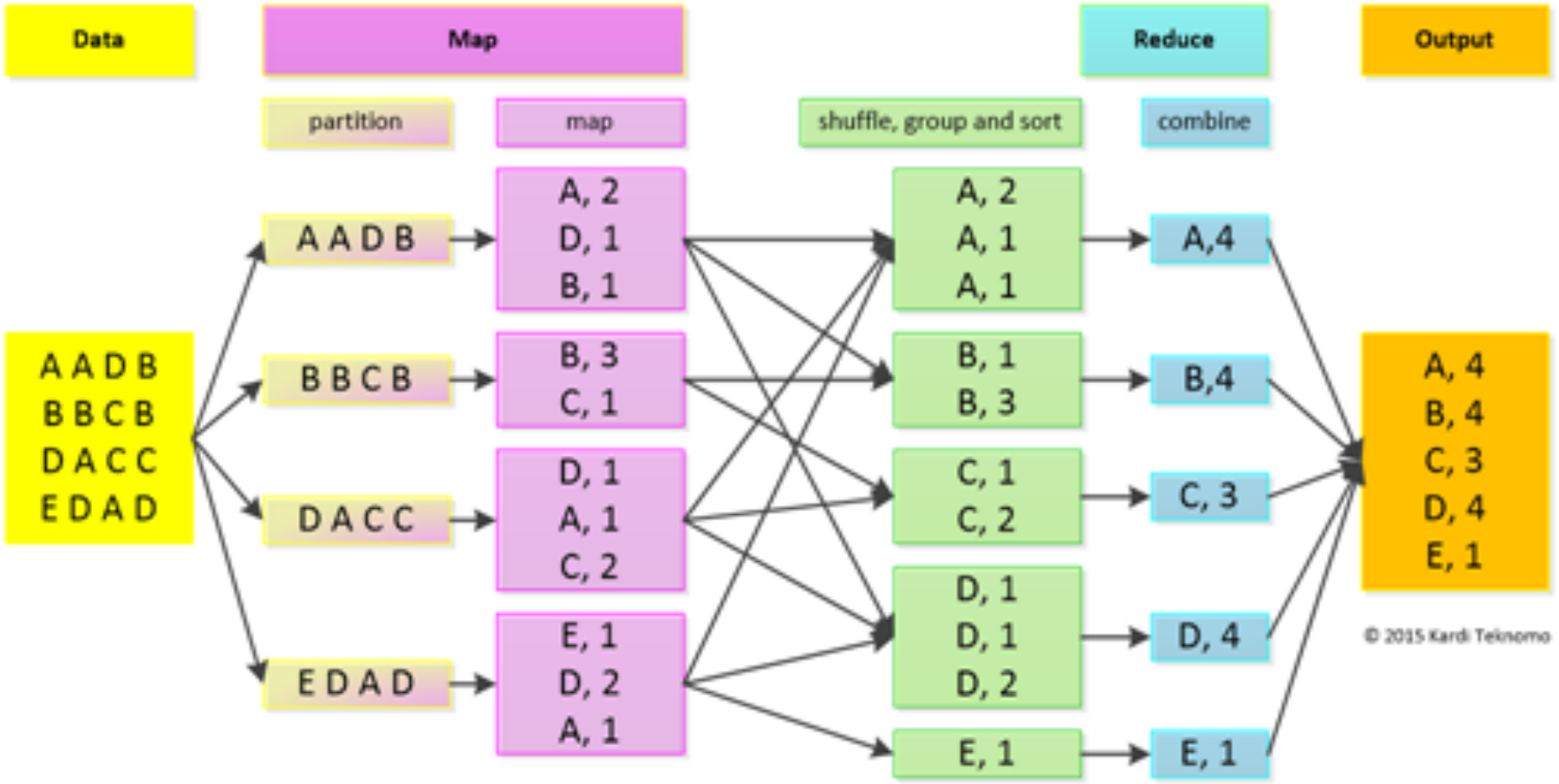
Google, Inc.

Abstract

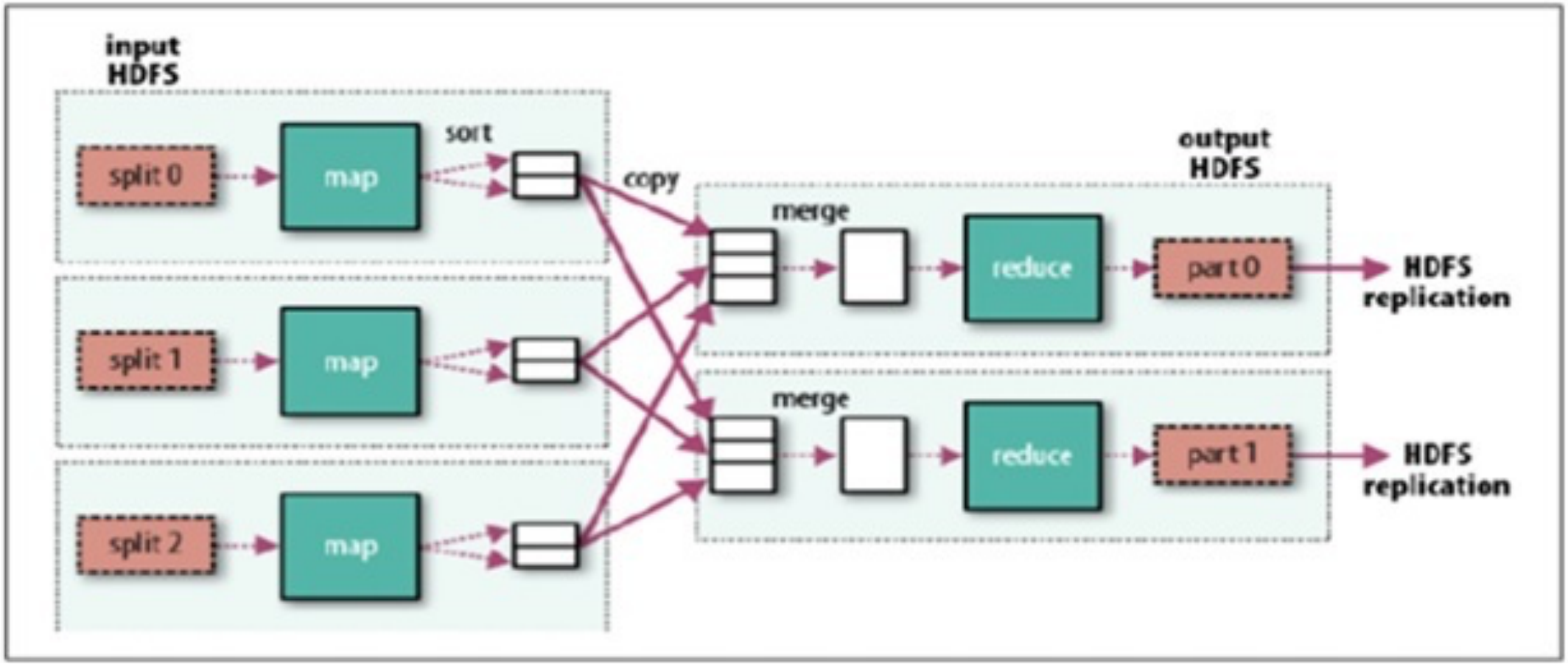
MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value

given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle

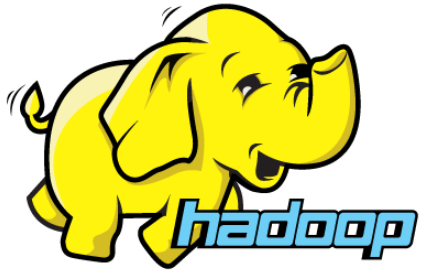
Map Reduce

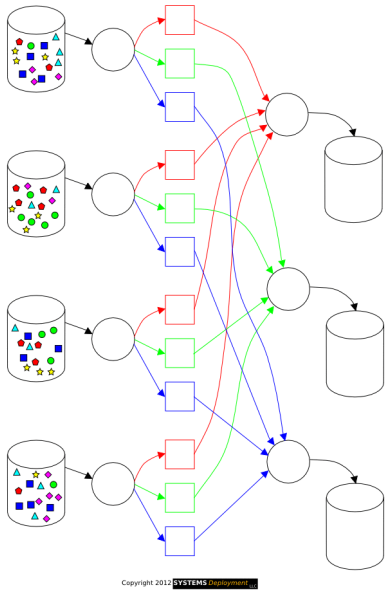

Data
Map
Reduce
Output
partition
map
shuffle, group and sort
combine
A A D B
B B C B
D A C C
E D A D
A A D B
B B C B
D A C C
E D A D
A, 2
D, 1
B, 1
B, 3
C, 1
D, 1
A, 1
C, 2
E, 1
D, 2
A, 1
A, 2
A, 1
A, 1
B, 1
B, 3
C, 1
C, 2
D, 1
D, 1
D, 2
E, 1
A,4
B,4
C, 3
D, 4
E, 1
A, 4
B, 4
C, 3
D, 4
E, 1
© 2015 Kardi Teknomo

Map Reduce on HDFS



Hadoop + Map Reduce



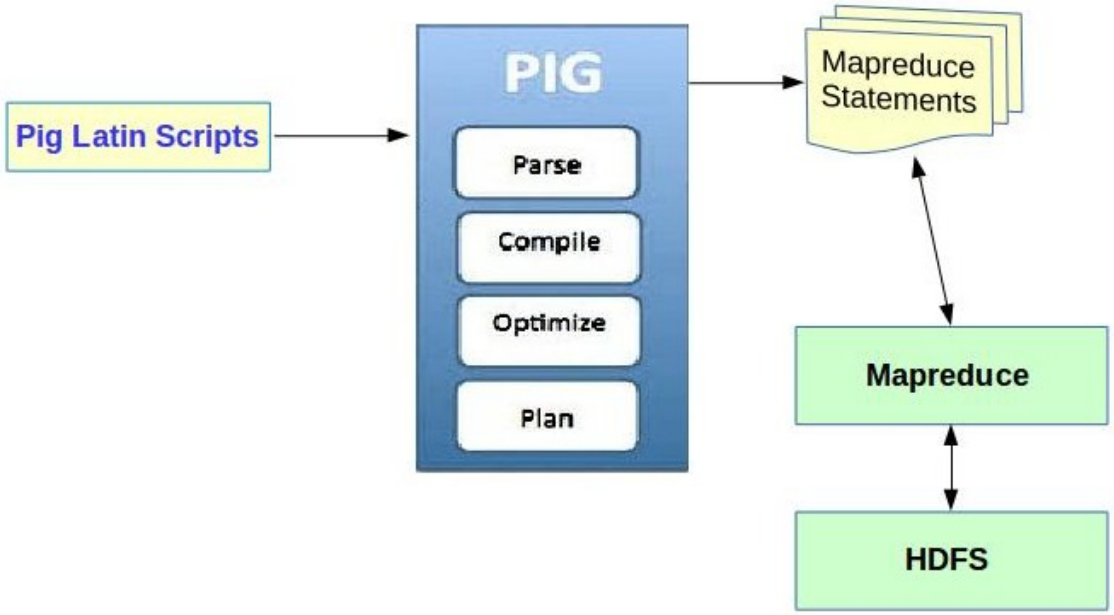


	RDBMS	Hadoop
Data size	Gigabytes *(terabytes)*	Petabytes *(hexabytes)*
Access	Interactive and batch	Batch
Updates	Read and write many times	Write once, read many times
Structure	Static schema	Dynamic schema
Integrity	High (ACID)	Low
Scaling	Nonlinear	Linear

Converting common and complex algorithms into MapReduce Framework involves many steps.
It is very slow and cumbersome

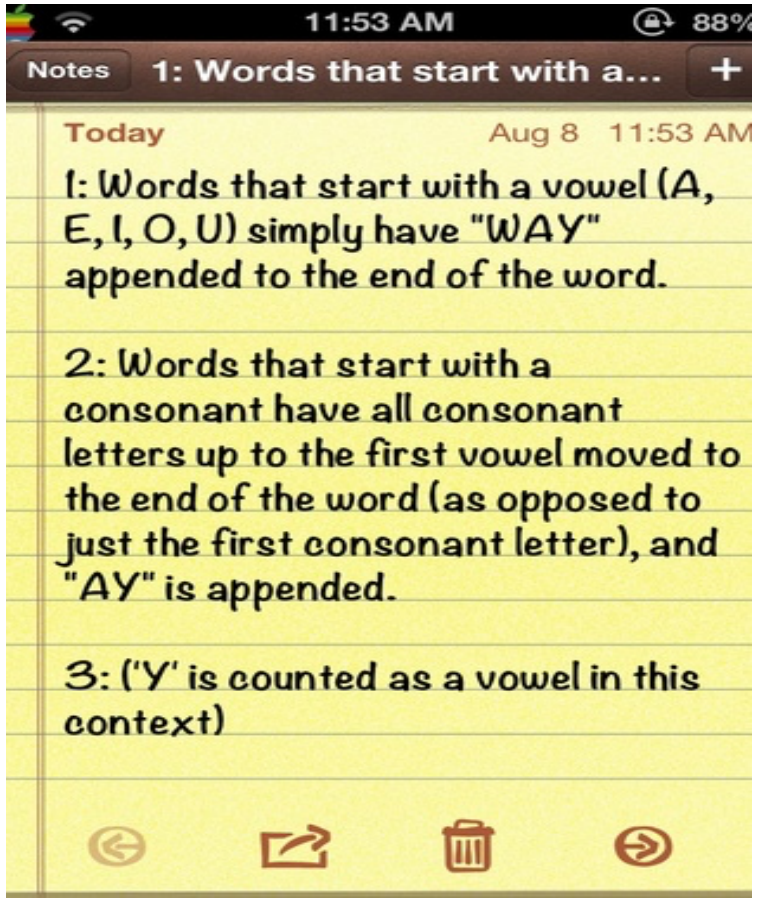
Pig

Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs.



Pig Latin

Apache Pig's Language layer is called Pig Latin.



Pig Latin

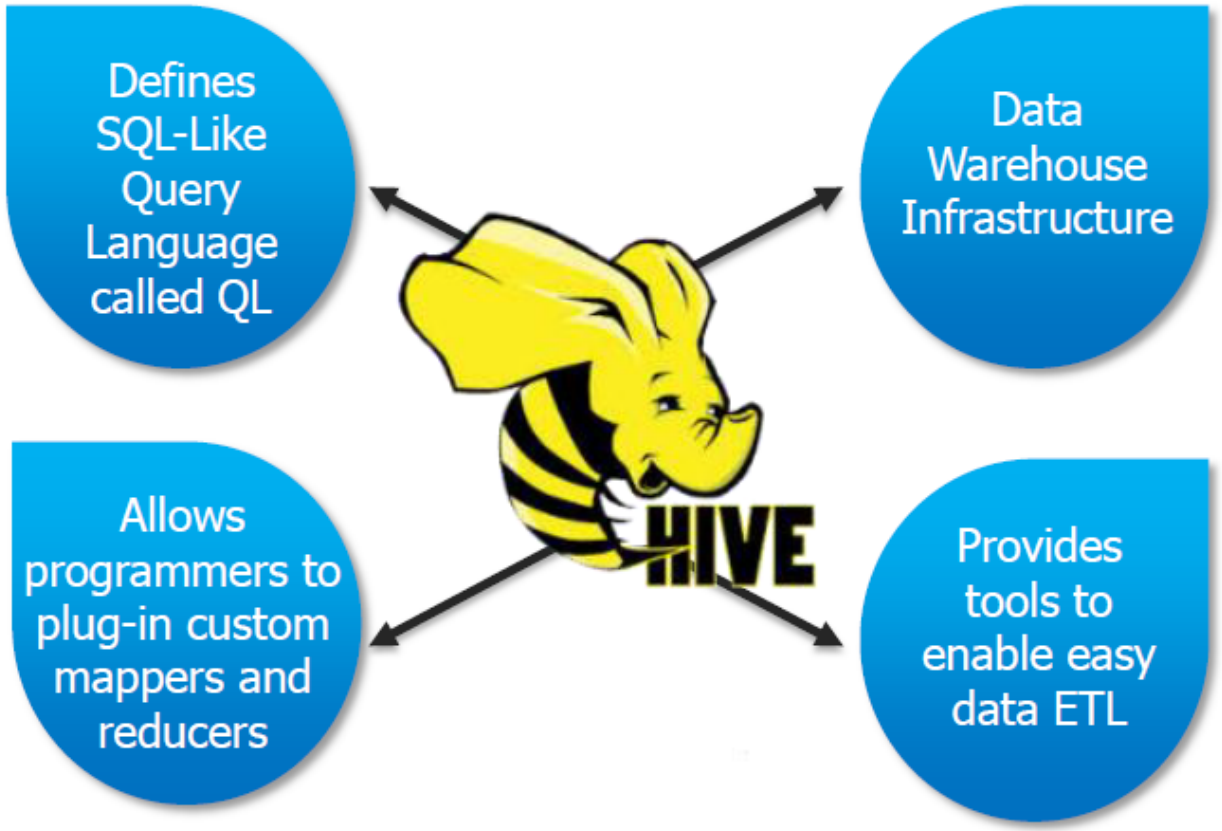
Pig Latin

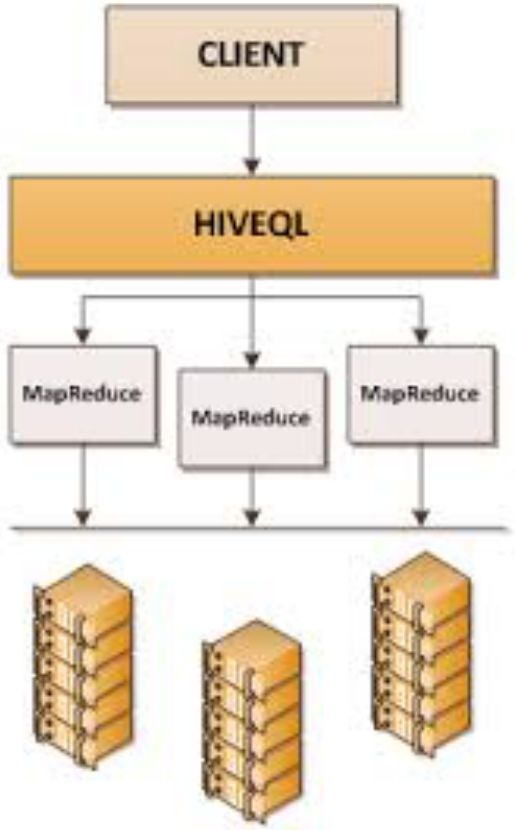
```
countrys = load '/user/gharriso/PIG_COUNTRIES' AS
 (country_id, country_name , country_subregion , region);

customers= load '/user/gharriso/PIG_CUSTOMERS' AS
  (cust_id,first_name, last_name, gender, yob, marital, postcode,city,country_id);

asianCountrys = filter countrys by region matches 'Asia';

joined = join customers by country_id, asianCountrys by country_id;

grouped = group joined by country_name;

agged = foreach grouped generate group, COUNT(joined.customers::cust_id);

morethan500cust = filter agged by $1 > 500;

ordered =order morethan500cust by $1 desc;

dump ordered;
```

- *No Java knowledge required*
- *Easy & Fast to write compared to Map Reduce*

- *Slow to run compared to Map Reduce because of translation*
- *New language to Master, but not intuitive like SQL*

HIVE


Defines SQL-Like Query Language called QL
Data Warehouse Infrastructure
Allows programmers to plug-in custom mappers and reducers
Provides tools to enable easy data ETL
HIVE


CLIENT
HIVEQL
MapReduce
MapReduce
MapReduce

HIVE Query Language

Pig Latin

```
countrys = load '/user/gharriso/PIG_COUNTRIES' AS
 (country_id, country_name , country_subregion , region);

customers= load '/user/gharriso/PIG_CUSTOMERS' AS
  (cust_id,first_name, last_name, gender, yob, marital, postcode,city,country_id);

asianCountrys = filter countrys by region matches 'Asia';

joined = join customers by country_id, asianCountrys by country_id;

grouped = group joined by country_name;

agged = foreach grouped generate group, COUNT(joined.customers::cust_id);

morethan500cust = filter agged by $1 > 500;

ordered =order morethan500cust by $1 desc;

dump ordered;
```

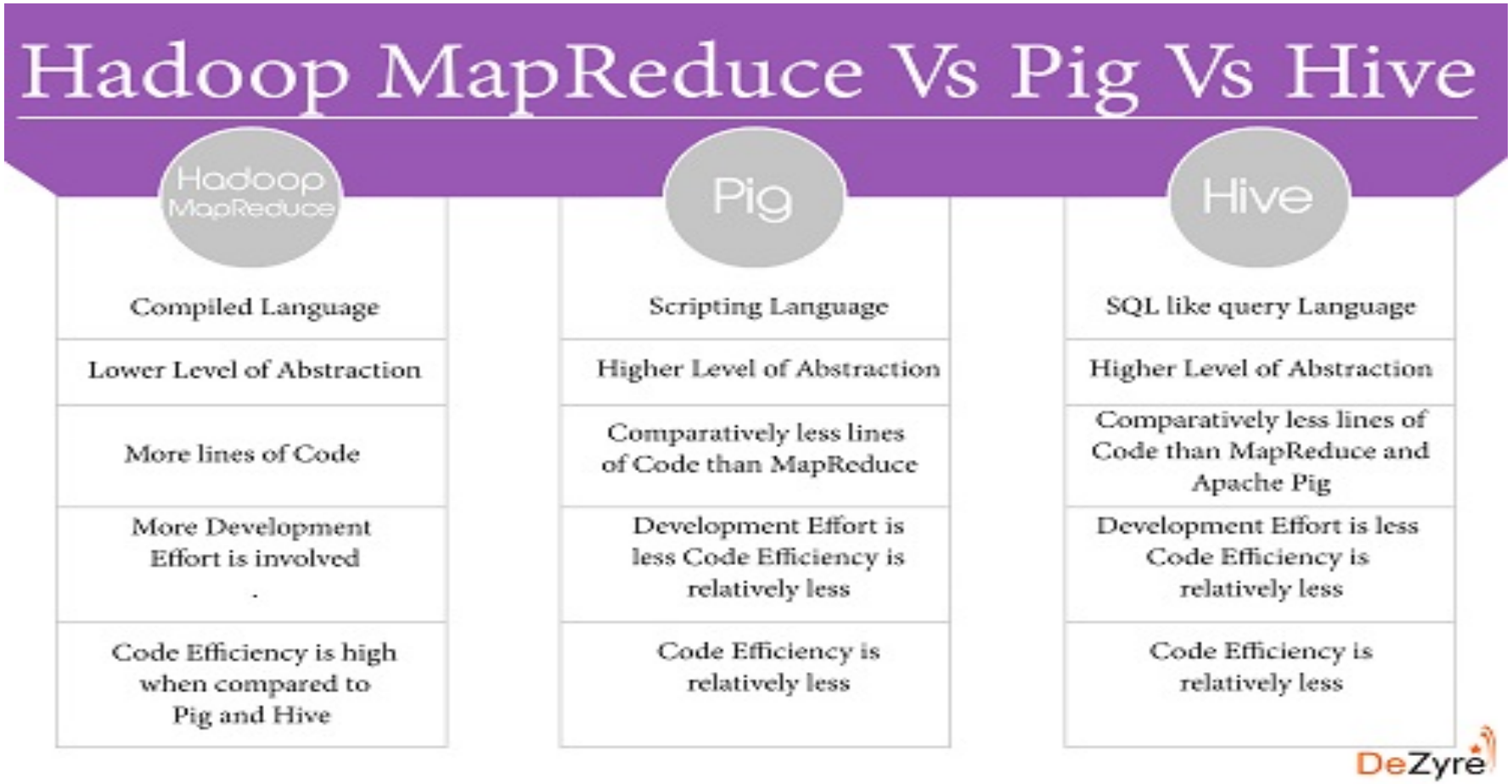
SQL or Hive QL

```
SELECT country_name,COUNT(cust_id) AS cust_count
   FROM countries co
  JOIN customers cu
    ON (co.country_id=cu.country_id)
  WHERE country_region='Asia'
  GROUP BY country_name
 HAVING COUNT(cust_id)>500
  ORDER BY cust_count DESC
```

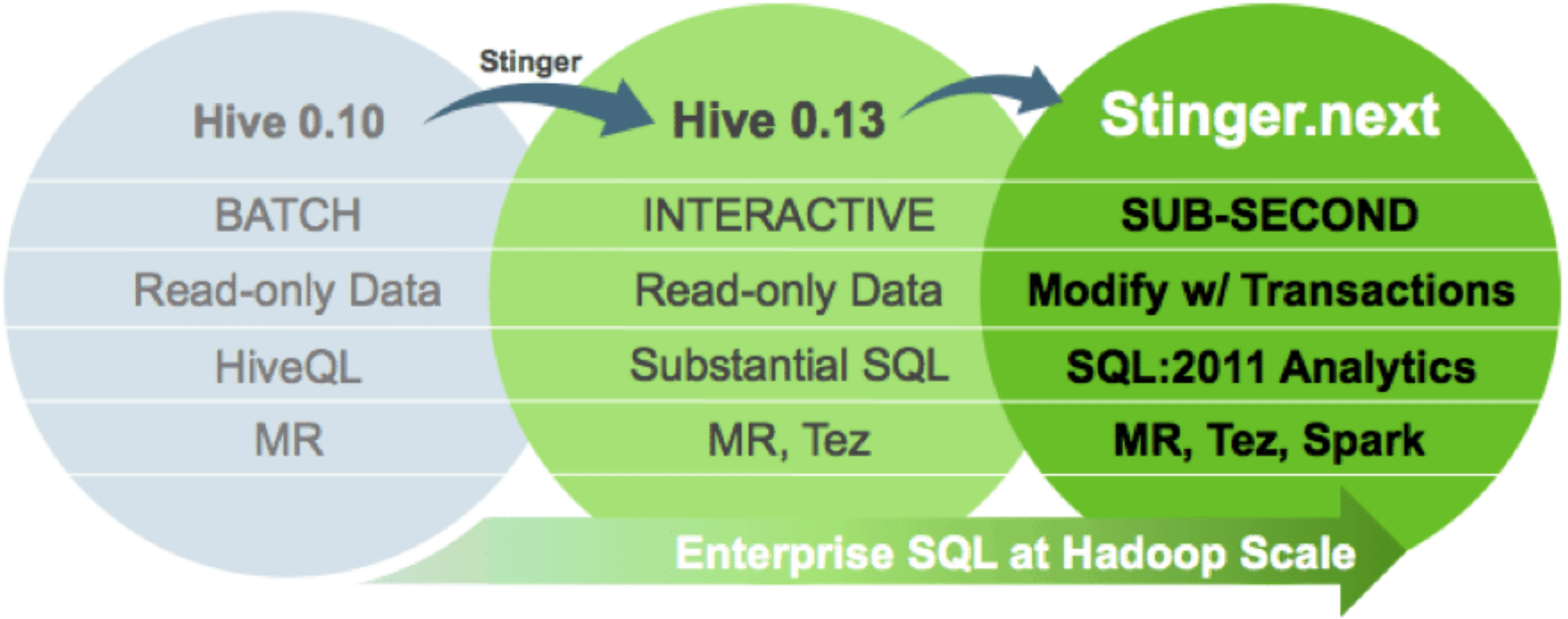
http://guyharrison.net

```
hive> desc emp;
OK
id                      int
name                    string
salary                  int
Time taken: 0.483 seconds, Fetched: 3 row(s)
hive> select * from emp;
OK
1       amit    80000
2       rakesh  35000
3       vishal  30000
4       rahul   50000
5       sameer  20000
1       ravi    30000
1       savi    26000
2       vikas   30000
```

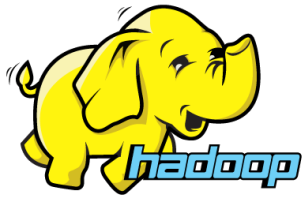
MapReduce vs Pig vs HIVE



HIVE


Stinger
Hive 0.10
BATCH
Read-only Data
HiveQL
MR
Hive 0.13
INTERACTIVE
Read-only Data
Substantial SQL
MR, Tez
Stinger.next
SUB-SECOND
Modify w/ Transactions
SQL:2011 Analytics
MR, Tez, Spark
Enterprise SQL at Hadoop Scale

Analysis



Batch Processing

How ?

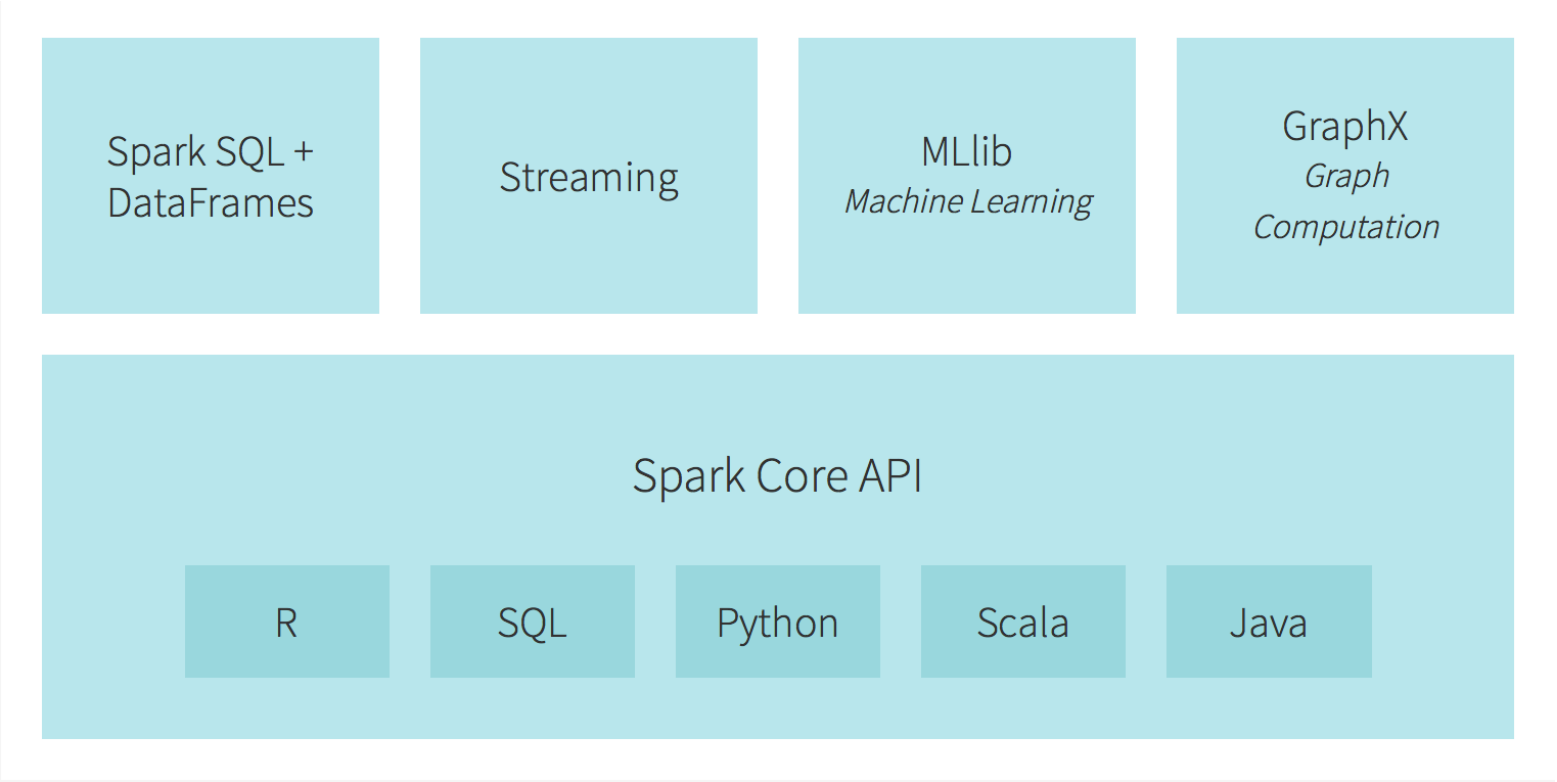
Real Time Processing

Interactive Analysis

Machine Learning

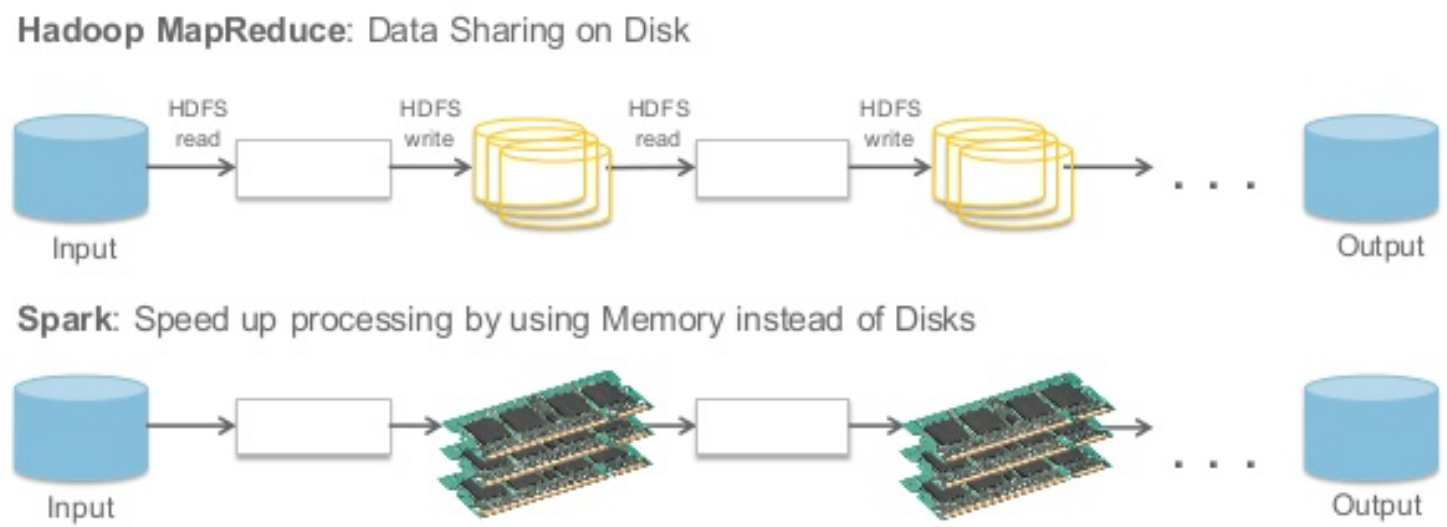
Graph Processing

Apache Spark

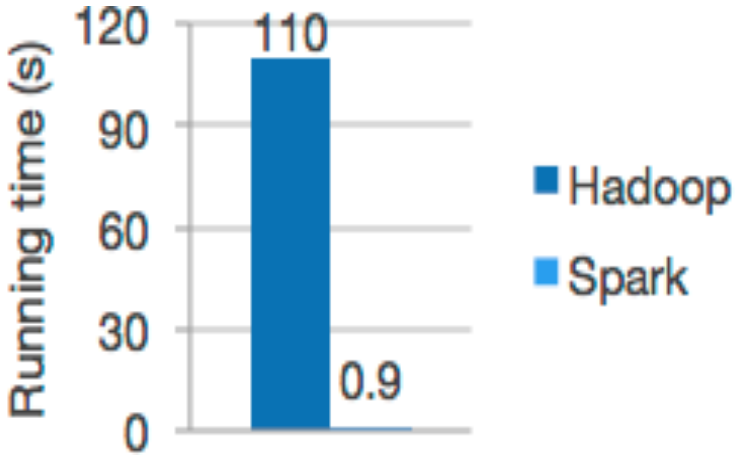


Apache Spark

Apache Spark utilizes in-memory caching and optimized execution for fast performance



Logistic Regression in Hadoop and Spark

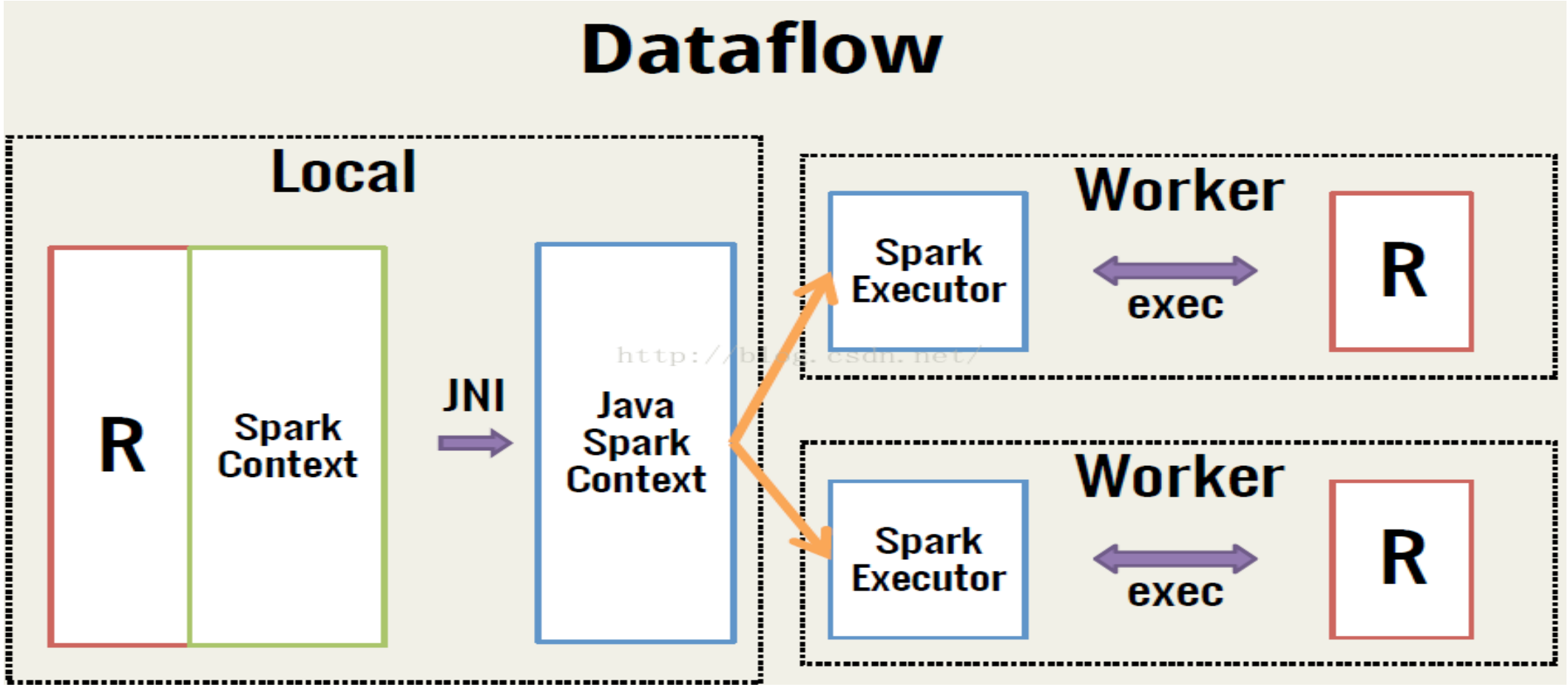


Apache Spark vs Hadoop MR

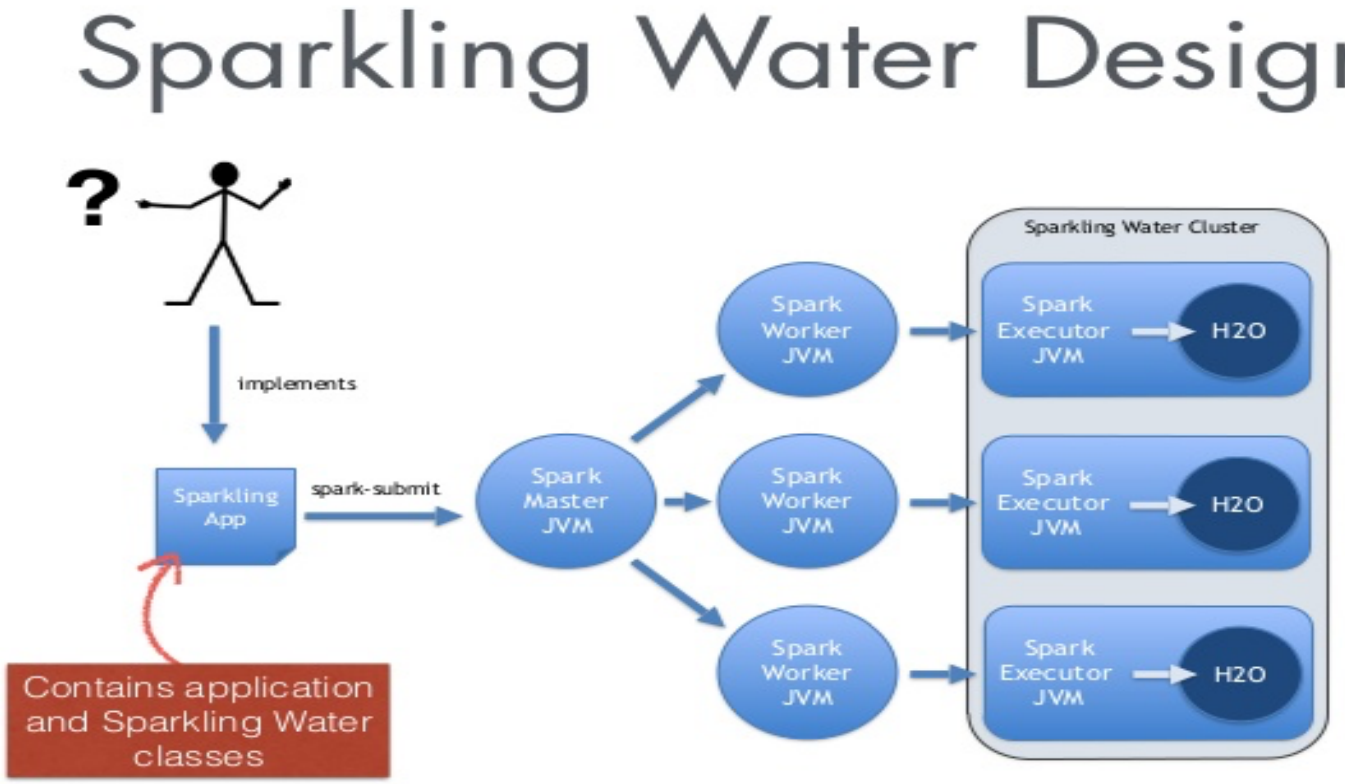
Parameters	Spark	Hadoop
Data Storage	Spark stores data in-memory.	Hadoop stores data on disk.
Fault tolerance	Spark's data storage model, resilient distributed datasets (RDD) guarantees fault tolerance.	It uses replication to achieve fault tolerance.
Line of code	Apache Spark is project of 20,000 Line of code.	Hadoop 2.0 has 1,20,000 Line of code
Speed	It is Faster due to In-memory computation.	It is relatively slower than Spark.
OS Support	• Linux • Windows • Mac OS	• Linux
High level language	• Scala • Python • Java • R	• Java
Streaming data	Spark can be used to process as well as modify real-time data with Spark streaming.	With Hadoop Map-Reduce one can process batch of stored data.
Machine Learning	Spark has its own set of Machine learning libraries (MLib).	Hadoop requires interface with other Machine learning library. Eg: Apache Mahout.

	Hadoop World Record	Spark 100 TB *	Spark 1 PB
Data Size	102.5 TB	100 TB	1000 TB
Elapsed Time	72 mins	23 mins	234 mins
# Nodes	2100	206	190
# Cores	50400	6592	6080
# Reducers	10,000	29,000	250,000
Rate	1.42 TB/min	4.27 TB/min	4.27 TB/min
Rate/node	0.67 GB/min	20.7 GB/min	22.5 GB/min
Sort Benchmark Daytona Rules	Yes	Yes	No
Environment	dedicated data center	EC2 (i2.8xlarge)	EC2 (i2.8xlarge)

Spark Dataflow in R

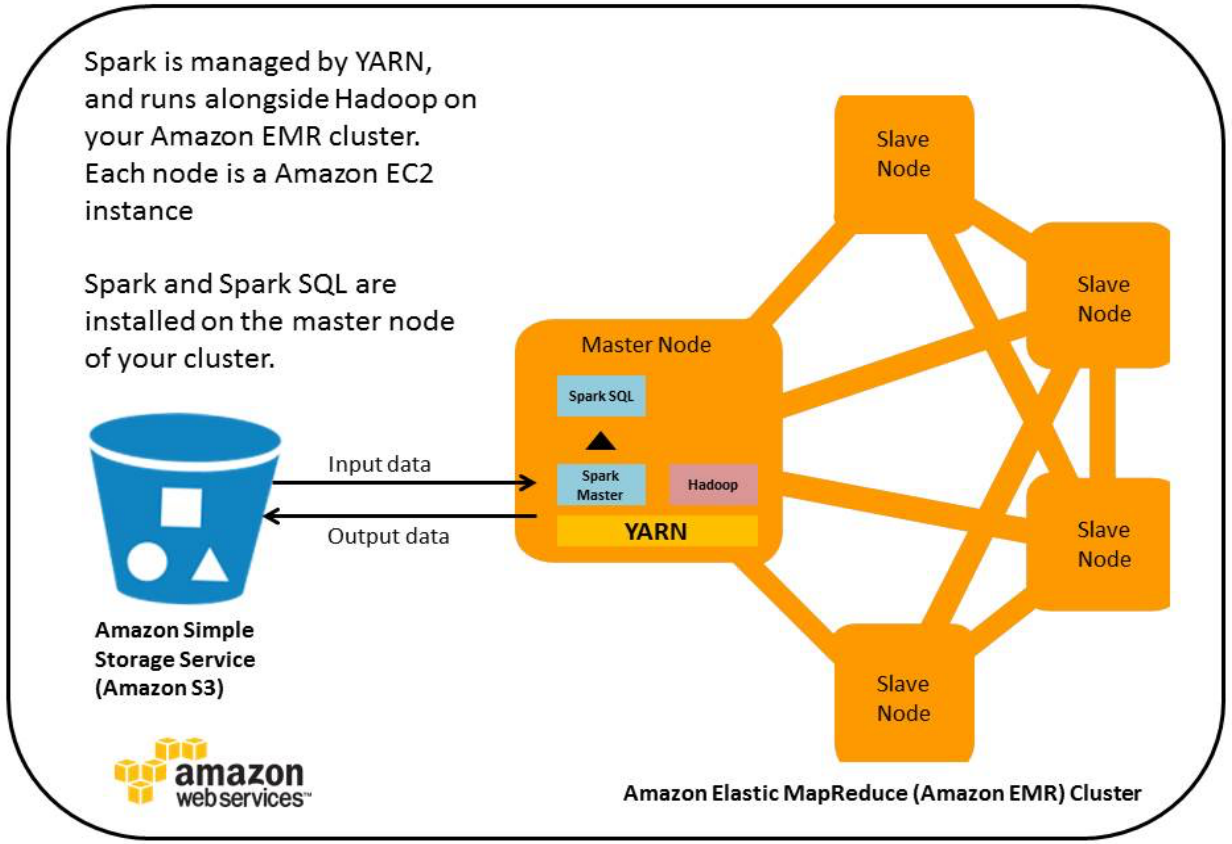


Spark + H2O = Sparkling Water

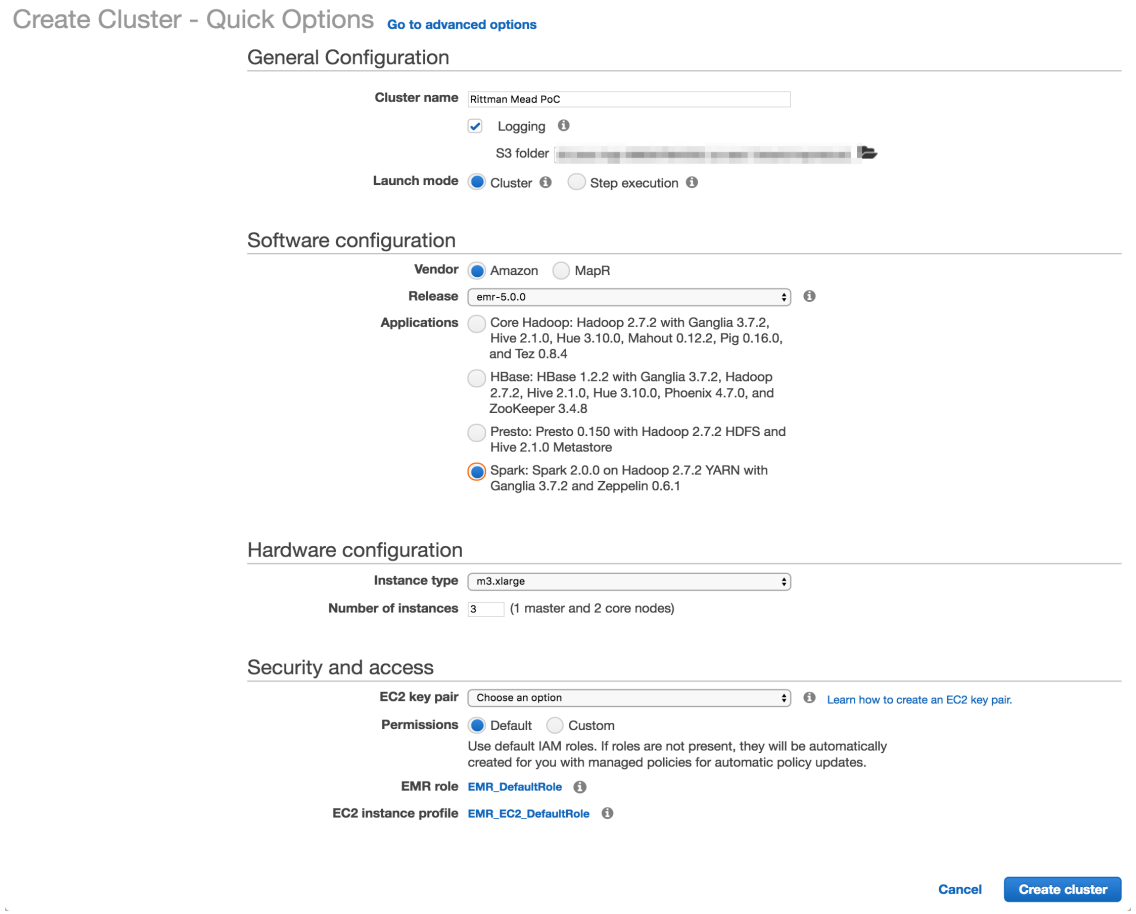


Distributed Machine Learning
on
Distributed Data Sets

Amazon Elastic Map Reduce (EMR)



Amazon Elastic Map Reduce (EMR)



Amazon Elastic Map Reduce (EMR)

Command Line Interface

Launch a 1 Master 5 Worker node m3.2x large cluster

With Hadoop, Hive, Spark, RStudio, Shiny, sparklyr all installed

```
aws emr create-cluster --applications Name=Hadoop Name=Spark Name=Hive Name=Pig Name=Tez Name=Ganglia
--release-label emr-5.2.0 --name "EMR 5.2.0 RStudio + sparklyr" --service-role EMR_DefaultRole \
--instance-groups InstanceGroupType=MASTER,InstanceCount=1,InstanceType=m3.2xlarge \
InstanceGroupType=CORE,InstanceCount=5,InstanceType=m3.2xlarge --bootstrap-actions \
Path=s3://aws-bigdata-blog/artifacts/aws-blog-emr-rstudio-sparklyr/rstudio_sparklyr_emr5.sh,\
Args=["--rstudio","--sparkr","--rexamples","--plyrmr","--rhdfs","--sparklyr"],\
Name="Install RStudio" --ec2-attributes InstanceProfile=EMR_EC2_DefaultRole,KeyName=<Your Key> \
--configurations '[{"Classification":"spark","Properties":{"maximizeResourceAllocation":"true"}}]' \
--region us-east-1
```

DSLA
DATA SCIENCE LEARNING ACADEMY

Break Time

To give AWS time to spin up the cluster …

Big Data

DSLA
DATA SCIENCE LEARNING ACADEMY

Student Presentations

While Flights data is being downloaded …

Demo

Flights using Spark on Amazon EMR

Data

DSLA
DATA SCIENCE LEARNING ACADEMY

Have good rest of weekend!