Databases in BIG Data

--From an application perspective

Dimensions of Big Data

Picture from https://www.techentice.com/the-data-veracity-big-data/
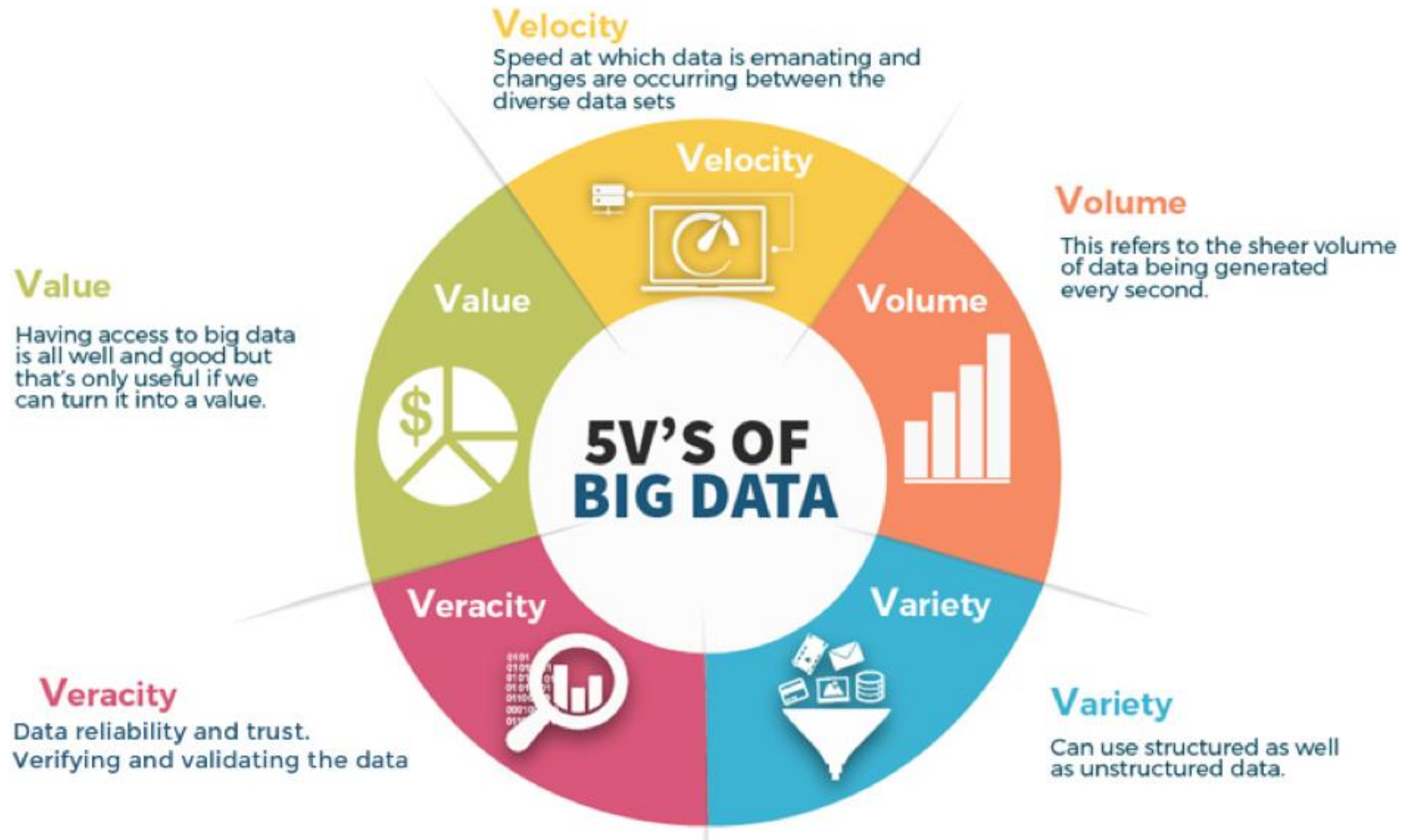
Organizing Big Data

- Database: A database is an **organized collection of data**, generally stored and accessed electronically from a computer system.

- DBMS: The database management system (DBMS) is the **software that interacts** with end users, applications, and the database itself to capture and analyze the data.

- Data ecosystem: A data ecosystem is a **collection of infrastructure**, analytics, and applications used to capture and analyze data. Data ecosystems provide companies with data that they rely on to understand their customers and to make better pricing, operations, and marketing decisions.

Definitions from https://www.wikipedia.org/

Different Databases (ecosystem)

MySQL (Relational Database)

Patient Contact Information

•	Patient ID	Last Name	First Name	Personal Email	Home Phone	Home Street	Home City	Home State	Home Zip
	1	Doe	Jane	jdoe@aol.com	555-1212	600 Table St	Sheetsville	VA	99999

MySQL (Relational Database)

Patient Contact Information

Patient ID	Last Name	First Name	Personal Email	Home Phone	Home Street	Home City	Home State	Home Zip
1	Doe	Jane	jdoe@aol.com	555-1212	600 Table St	Sheetsville	VA	99999
2	Deer	James	jdeer@mindspring.net	555-2121	800 Relation Drive	Jointown	NC	98989

MySQL (Relational Database)

Patient Contact Information

	Patient ID	Last Name	First Name	Personal Email	Home Phone	Work Phone	Cell Phone	Home Street	Home City	Home State	Home Zip
	1	Doe	Jane	jdoe@aol.com	555-1212		555-9090	600 Table St	Sheetsville	VA	99999
	2	Deer	James	jdeer@mindspring.net	555-2121	555-1111		800 Relation Drive	Jointown	NC	98989

MySQL (Relational Database)

Patient Contact Information

Patient ID	Last Name	First Name	Personal Email	Work Email	Home Phone	Work Phone	Cell Phone	Emergency Contact Phone	Home Street	Home City	Home State	Home Zip	Work Street	Work City	Work State	Work Zip
1	Doe	Jane	jdoe@aol.com		555-1212		555-9090		600 Table St	Sheetsville	VA	99999	700 Column Row	Sheetsville	VA	99999
2	Deer	James	jdeer@mindspring.net		555-2121	555-1111			800 Relation Drive	Jointown	NC	98989				
3	Apryl	Lenser	jorgb@sbcglobal.net		202-555-0116		202-555-0193		676 George Street	Sheboygan	WI	53081				
4	Delmy	Trammel	johndo@icloud.com		202-555-0107	202-555-0107	202-555-0195		69 Coffee Dr.	Ronkonkoma	NY	11779				
5	Sharolyn	Spence	carmena@comcast.net	jorgb@sbcglobal.net	202-555-0165			202-555-0183	8778 Spruce Ave.	Stratford	CT	6614				
6	Cassondra	Yeats	helger@mac.com		202-555-0109		202-555-0124		463 SmokyHollow St	Carmel	NY	10512				
7	Margo	Varney	amichalo@mac.com		202-555-0192		202-555-0188		42 S. Highland Lane	Hartselle	AL	35640				
8	Mira	Pfaff	rasca@yahoo.com		202-555-0198		202-555-0168		492 Pearl Street	Saint Charles	IL	60174				
9	Raymon	Ryer	ghost@gmail.com		202-555-0150	202-555-0150			21 St Louis Street	Clarksville	TN	37040	3 SE. Logan St.	Arvada	CO	80003
10	Kimbra	Gravel	parrt@aol.com		202-555-0142				7576 Homewood St	Davison	MI	48423				
11	Sumiko	Cullinan	padme@yahoo.ca		202-555-0154		202-555-0170		31 Nichols Court	Nanuet	NY	10954				
12	Boris	Ellzey	schwaang@verizon.net	johndo@icloud.com	202-555-0182				8224 Eagle Drive	Rome	NY	13440				
13	Dagmar	Morano	dougj@yahoo.com		202-555-0126				3 Grove Dr.	Saint Cloud	MN	56301				
14	Trista	Knuckles	anicolao@verizon.net		614-555-0184		614-555-0188		370 Big Rock Cove	Olive Branch	MS	38654				
15	Lenny	Walcott	harpes@optionline.net		614-555-0119				2 Philmont Avenue	Lapeer	MI	48446	560 W Hudson St.	Olympia	WA	98512
16	Jodie	Manion	mpiotr@comcast.net		614-555-0125			614-555-0519	960 Sycamore St.	Enfield	CT	6082				
17	Erlinda	Eisenmenger	twoflower@msn.com		614-555-0118	614-555-0118			77 Hillcrest Court	Eden Prairie	MN	55347				
18	Jaquelyn	Daffron	mwilson@msn.com		614-555-0196		614-555-0128		94 Charles Rd.	Williamstown	NJ	8094				
19	Barbie	Brandis	carreras@outlook.com		614-555-0187				11 Grove Drive	Sioux Falls	SD	57103				
20	Apryl	Lenser	jorgb@sbcglobal.net	apryl@comcast.net	202-555-0116				676 George Street	Sheboygan	WI	53081				

MySQL (Relational Database)

Patient Contact Information

Patient ID	Last Name	First Name	Personal Email	Work Email	Home Phone	Work Phone	Cell Phone	Emergency Contact Phone	Home Street	Home City	Home State	Home Zip	Work Street	Work City	Work State	Work Zip
1	Doe	Jane	jdoe@aol.com		555-1212		555-9090		600 Table St	Sheetsville	VA	99999	700 Column Row	Sheetsville	VA	99999
2	Deer	James	jdeer@mindspring.net		555-2121	555-1111			800 Relation Drive	Jointown	NC	98989				
3	Apryl	Lenser	jorgb@sbcglobal.net		202-555-0116		202-555-0193		676 George Street	Sheboygan	WI	53081				
4	Delmy	Trammel	johndo@icloud.com		202-555-0107	202-555-0107	202-555-0195		69 Coffee Dr.	Ronkonkoma	NY	11779				
5	Sharolyn	Spence	carmena@comcast.net	jorgb@sbcglobal.net	202-555-0165			202-555-0183	8778 Spruce Ave.	Stratford	CT	6614				
6	Cassondra	Yeats	helger@mac.com		202-555-0109		202-555-0124		463 SmokyHollow St	Carmel	NY	10512				
7	Margo	Varney	amichalo@mac.com		202-555-0192		202-555-0188		42 S. Highland Lane	Hartselle	AL	35640				
8	Mira	Pfaff	rasca@yahoo.com		202-555-0198		202-555-0168		492 Pearl Street	Saint Charles	IL	60174				
9	Raymon	Ryer	ghost@gmail.com		202-555-0150	202-555-0150			21 St Louis Street	Clarksville	TN	37040	3 SE. Logan St.	Arvada	CO	80003
10	Kimbra	Gravel	parrt@aol.com		202-555-0142				7576 Homewood St	Davison	MI	48423				
11	Sumiko	Cullinan	padme@yahoo.ca		202-555-0154		202-555-0170		31 Nichols Court	Nanuet	NY	10954				
12	Boris	Ellzey	schwaang@verizon.net	johndo@icloud.com	202-555-0182				8224 Eagle Drive	Rome	NY	13440				
13	Dagmar	Morano	dougj@yahoo.com		202-555-0126				3 Grove Dr.	Saint Cloud	MN	56301				
14	Trista	Knuckles	anicolao@verizon.net		614-555-0184		614-555-0188		370 Big Rock Cove	Olive Branch	MS	38654				
15	Lenny	Walcott	harpes@optionline.net		614-555-0119				2 Philmont Avenue	Lapeer	MI	48446	560 W Hudson St.	Olympia	WA	98512
16	Jodie	Manion	mpiotr@comcast.net		614-555-0125			614-555-0519	960 Sycamore St.	Enfield	CT	6082				
17	Erlinda	Eisenmenger	twoflower@msn.com		614-555-0118	614-555-0118			77 Hillcrest Court	Eden Prairie	MN	55347				
18	Jaquelyn	Daffron	mwilson@msn.com		614-555-0196		614-555-0128		94 Charles Rd.	Williamstown	NJ	8094				
19	Barbie	Brandis	carreras@outlook.com		614-555-0187				11 Grove Drive	Sioux Falls	SD	57103				
20	Apryl	Lenser	jorgb@sbcglobal.net	apryl@comcast.net	202-555-0116				676 George Street	Sheboygan	WI	53081				

Pictures from MangoDB https://www.youtube.com/watch?v=EE8ZTQxa0AM

MySQL (Relational Database)

Patient Phone Numbers

	Patient ID	Phone Number	Label
	1	555-1212	Home
	1	555-9090	Cell
	2	555-2121	Home
	2	555-1111	Work

Patient Names

	Patient ID	Last Name	First Name
	1	Doe	Jane
	2	Deer	James

Patient Addresses

	Patient ID	Street	City	State	Zip	Label
	1	600 Table St	Sheetsville	VA	99999	Home
	1	700 Column Row	Sheetsville	VA	99999	Work
	2	800 Relation Drive	Jointown	NC	98989	Home

Lab Results

	Patient ID	Test ID	Test Date	Results
	1	46	1/15/2017	7/18
	1	35564	8/23/2016	80mg
	2	2643	8/23/2016	negative

Patient Emergency Contacts

	Patient ID	Last Name	First Name	Phone	Relationship
	1	Doe	Jon	xxx-xxxx	Brother
	2	Deer	Cynthia	xxx-xxxx	Mother
	3	Sprat	Jill	xxx-xxxx	Spouse

Patient Meds

	Patient ID	Date Prescribed	Valid Through	Medication Name	Dosage	Notes	Physician ID
	1	1/15/2017	2/15/2017	Hyberdoze	15mg 2x daily		245-23415-13
	1	9/1/2015	11/15/2015	Comanapracil	6mg 1x daily	take with food	245-23415-13
	1	8/23/2016	9/14/2016	Vaxadrin	250mg 4x		832-82356-89

Patient Visits

	Patient ID	Date	Physician ID	Reason for Visit	Notes		Label
	1	4/19/2014	245-23415-13	mysterious tingling	patient had a battery on their tongue		Home
	2	4/2/16	832-82356-89	numbness in third arm	no third arm actually present		Work
	3	11/9	456-83223-11	double vision	sounds fun!		Home

MongoDB

Pictures from MangoDB https://www.youtube.com/watch?v=EE8ZTQxa0AM

MongoDB

Pictures from MangoDB https://www.youtube.com/watch?v=EE8ZTQxa0AM
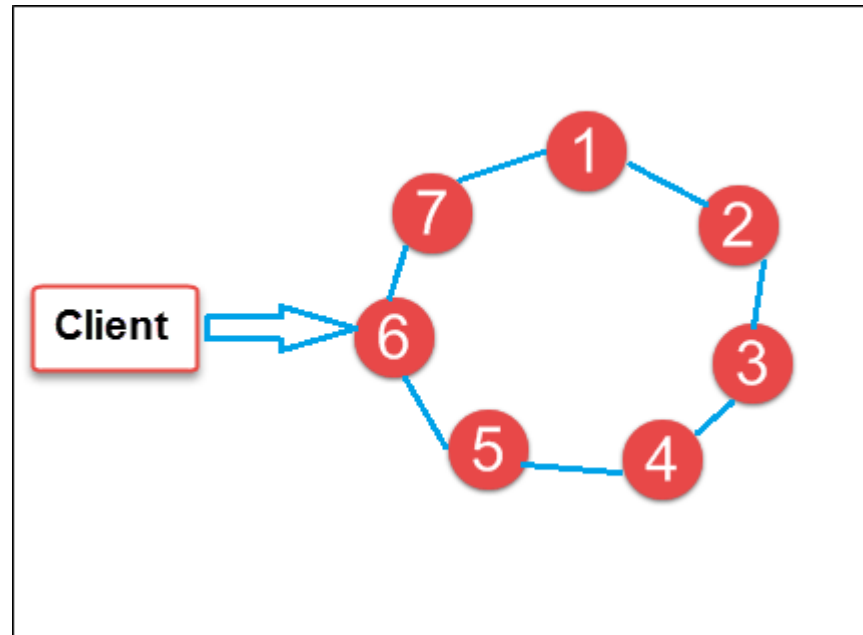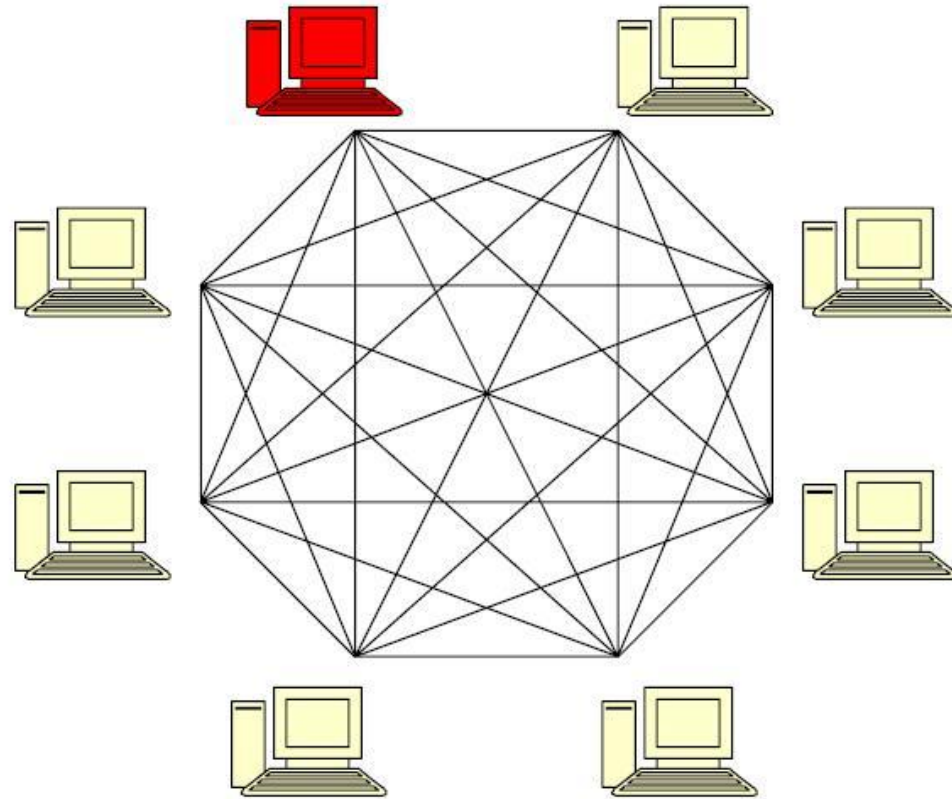
Application

- Lots of data
  - Copies of messages, reverse indices of messages, per user data.
- Many incoming requests resulting in a lot of **random** reads and **random** writes.

Cassandra

- Lots of data
  - Copies of messages, reverse indices of messages, per user data.
- Many incoming requests resulting in a lot of **random** reads and **random** writes.
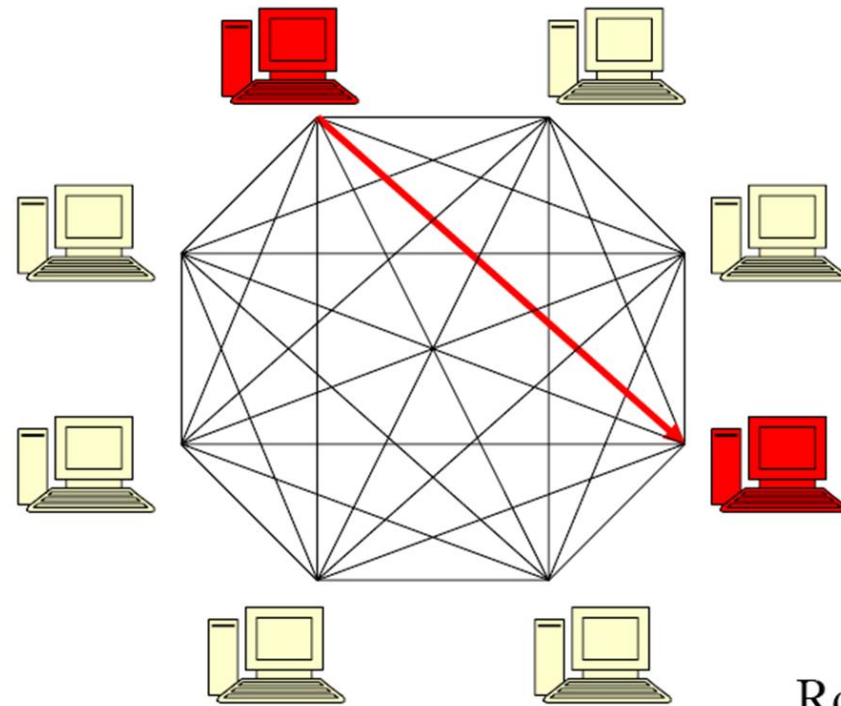
Cassandra

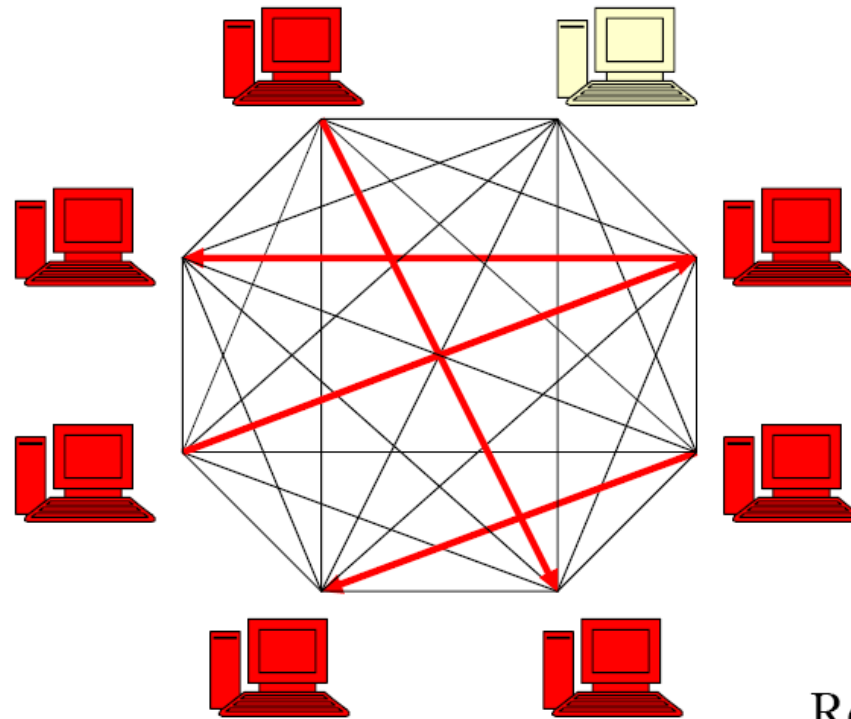Pictures from https://www.guru99.com/cassandra-tutorial.html

Cassandra

Gossip

Adopted from slides by By Perry Hoekstra, Jiaheng Lu, Avinash Lakshman, Prashant Malik, and Jimmy Lin

Cassandra

Gossip

Round 1: 2

Adopted from slides by By Perry Hoekstra, Jiaheng Lu, Avinash Lakshman, Prashant Malik, and Jimmy Lin

Cassandra

Cassandra

- Lots of data
  - Copies of messages, reverse indices of messages, per user data.
- Many incoming and time sensitive requests resulting in a lot of **random** reads and **random** writes.

Some applications

- **Lots of Lots** of data
  - Copies of messages, reverse indices of messages, per user data.
- **Many many** incoming data and **time in-sensitive** requests resulting in a lot of random reads and random writes.
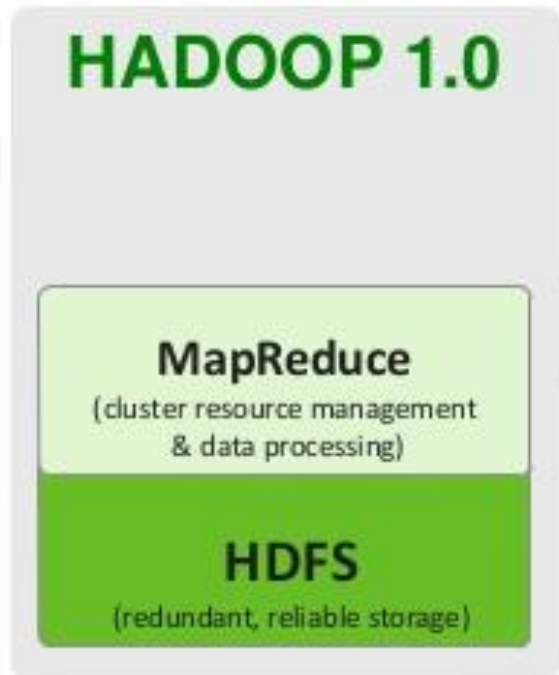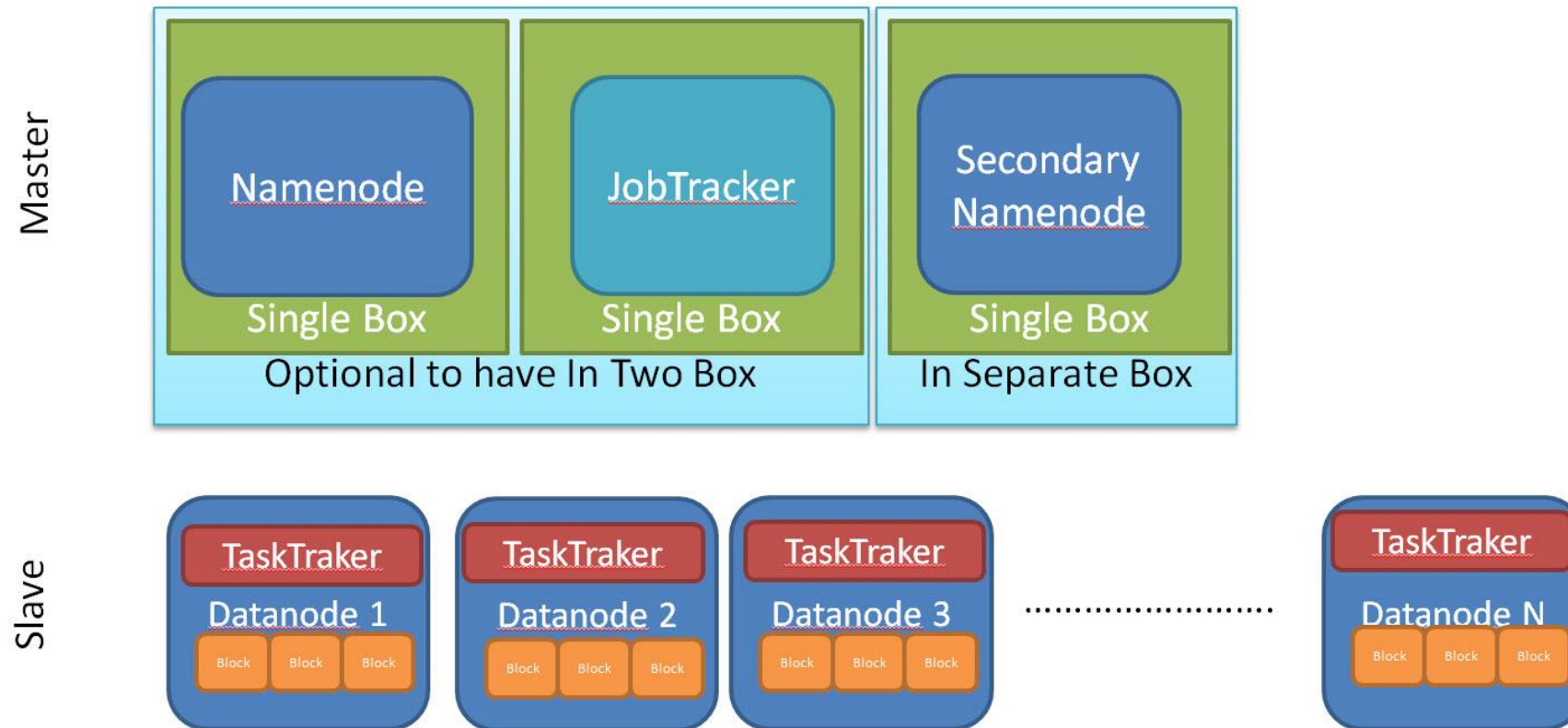
Hadoop

- **Lots of Lots** of data
  - Copies of messages, reverse indices of messages, per user data.
- **Many many** incoming data and **time in-sensitive** requests resulting in a lot of random reads and random writes.

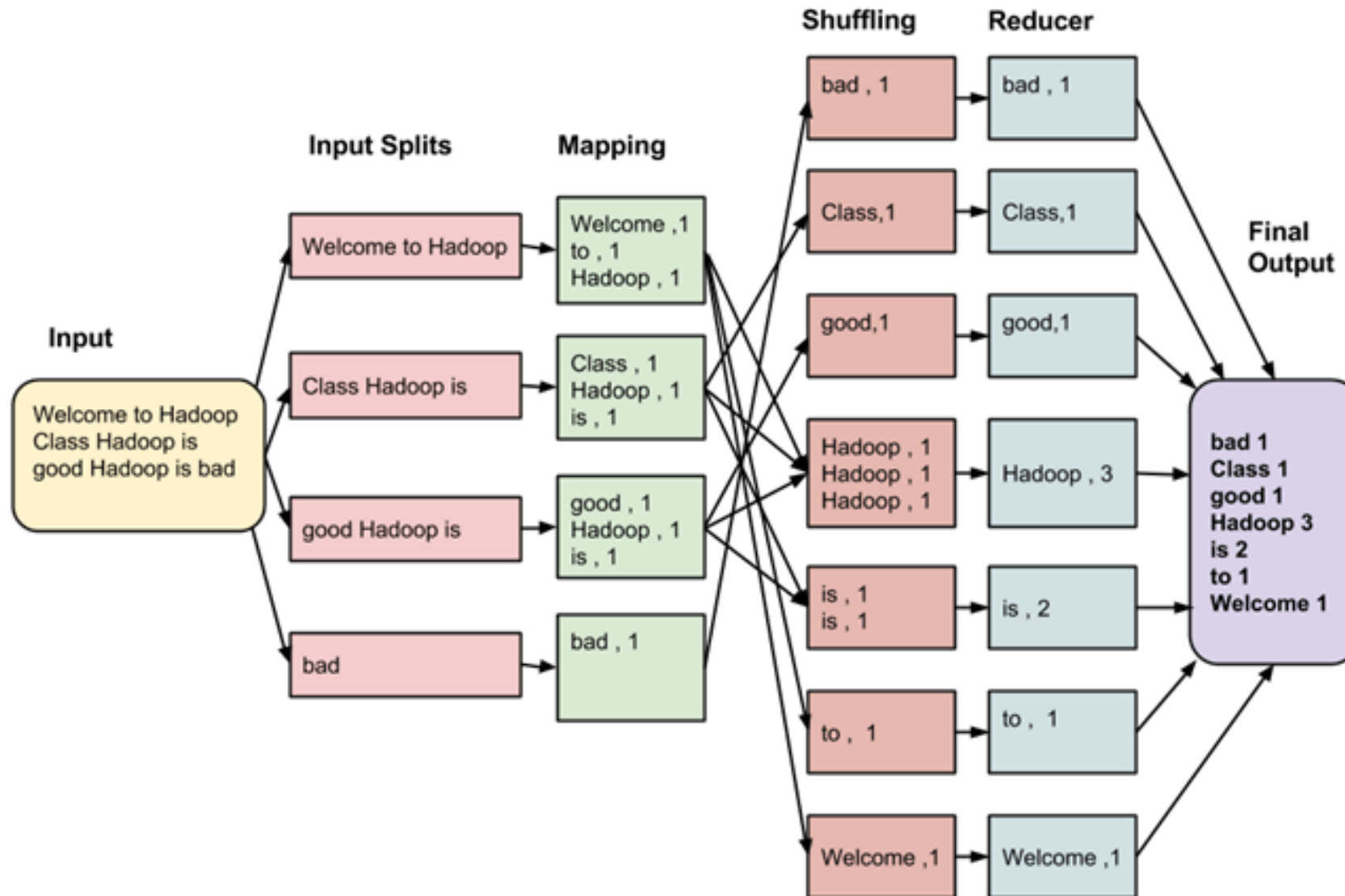# Hadoop

Single Use System
Batch Apps

HADOOP 1.0

MapReduce
(cluster resource management & data processing)

HDFS
(redundant, reliable storage)

Hadoop

Master

Namenode	JobTracker	Secondary Namenode
Single Box	Single Box	Single Box
Optional to have In Two Box		In Separate Box

Slave

TaskTraker	TaskTraker	TaskTraker	....	TaskTraker
Datanode 1	Datanode 2	Datanode 3		Datanode N
Block Block Block	Block Block Block	Block Block Block		Block Block Block

Picture from https://blogs.msdn.microsoft.com/avkashchauhan/2012/02/24/master-slave-architecture-in-hadoop/
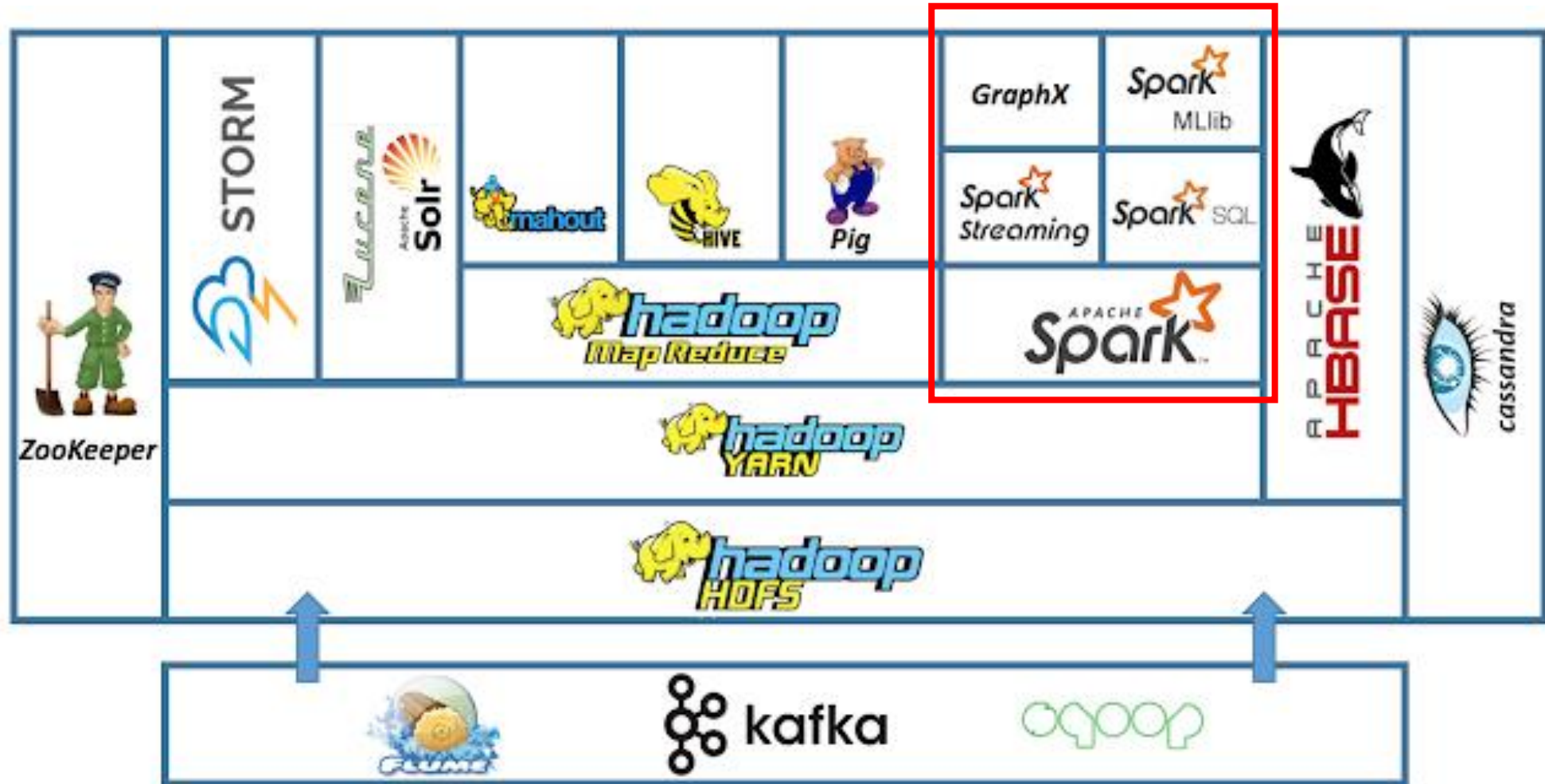
Hadoop

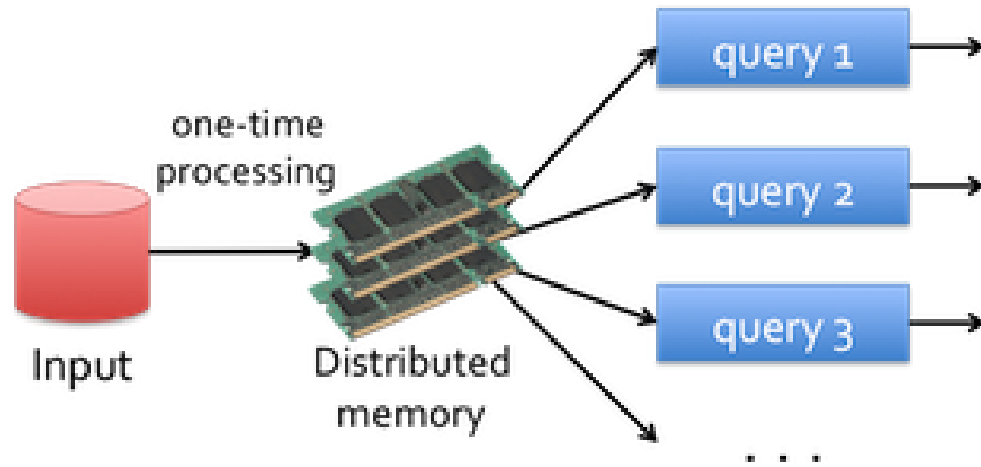Picture from https://www.guru99.com/introduction-to-mapreduce.html

Hadoop

Pictures from https://www.hdfstutorial.com/blog/hadoop-1-vs-hadoop-2-differences/

Hadoop Ecosystem

Picture from http://blog.newtechways.com/2017/10/apache-hadoop-ecosystem.html
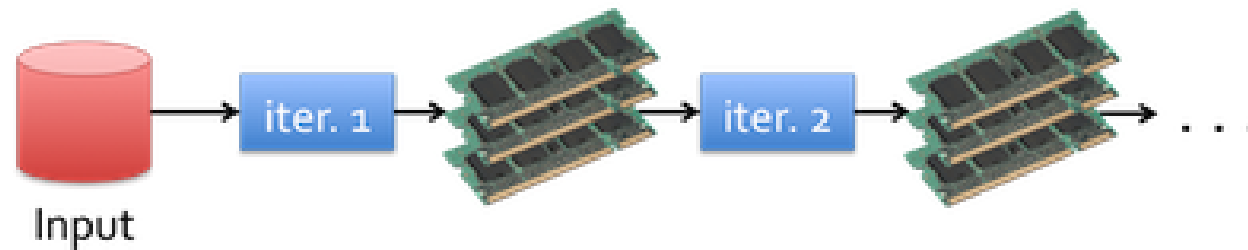
Spark

(a) Low-latency computations (queries)

(b) Iterative computations

Picture from https://databricks.com/blog/2013/11/21/putting-spark-to-use.html

Spark Ecosystem

Picture from https://data-flair.training/blogs/apache-spark-ecosystem-components/
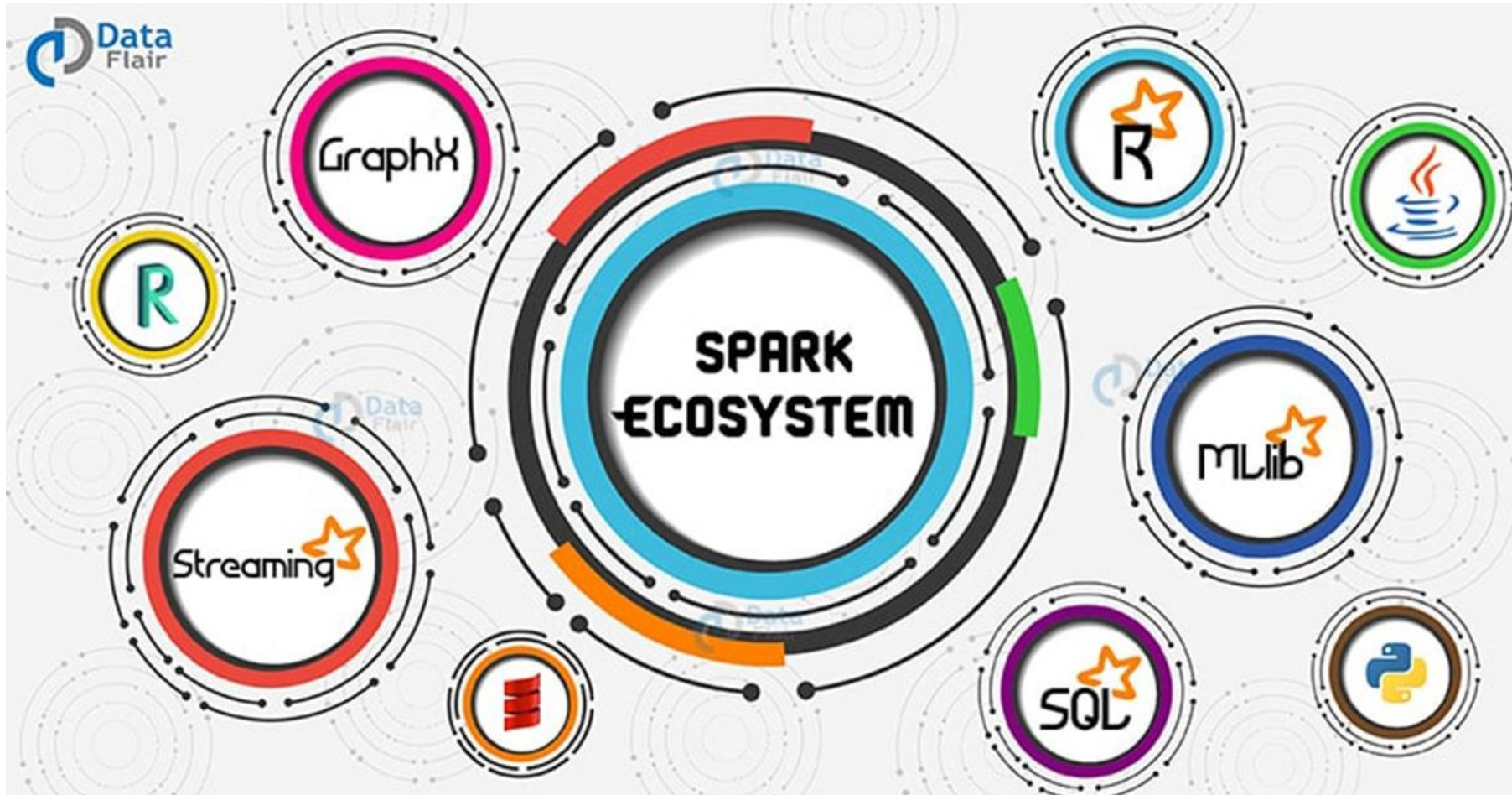
Comparison (Hadoop vs. SQL)

- Data Size

Picture from https://www.educba.com/hadoop-vs-sql/
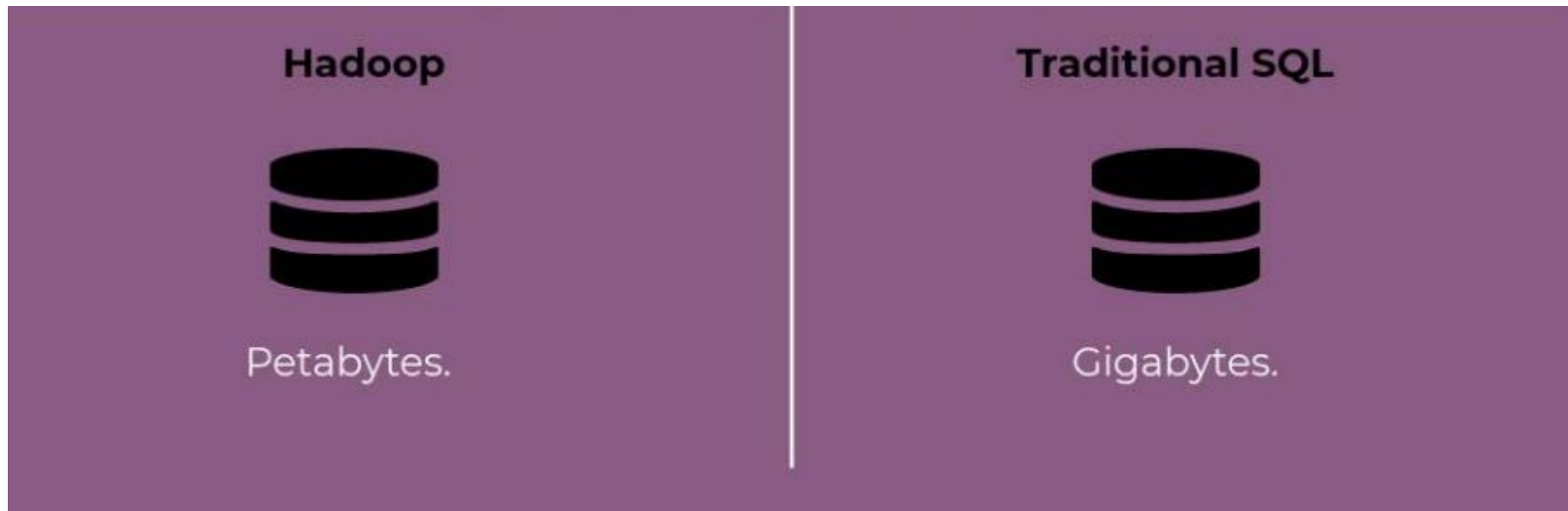
Comparison (Hadoop vs. SQL)

- Scaling

Picture from https://www.educba.com/hadoop-vs-sql/

Comparison (Hadoop vs. SQL)

- Query Speed

INSERT, SELECT type statements are very fast in Hadoop compared to SQL.

SQL syntax are much slower when executed on millions of rows at a time.

Hadoop

SQL

Picture from https://www.educba.com/sql-vs-hadoop/

MySQL (Relational Database)

Patient Contact Information

	Patient ID	Last Name	First Name	Personal Email	Home Phone	Work Phone	Cell Phone	Home Street	Home City	Home State	Home Zip
	1	Doe	Jane	jdoe@aol.com	555-1212		555-9090	600 Table St	Sheetsville	VA	99999
	2	Deer	James	jdeer@mindspring.net	555-2121	555-1111		800 Relation Drive	Jointown	NC	98989

Comparison (Hadoop vs. SQL)

- Query Speed

INSERT, SELECT type statements are very fast in Hadoop compared to SQL.

Hadoop

SQL syntax are much slower when executed on millions of rows at a time.

SQL

Picture from https://www.educba.com/sql-vs-hadoop/

Comparison (Hadoop vs. SQL)

- Machine Learning Support

Advanced machine learning and artificial intelligence techniques can be build using Hadoop.

Hadoop

Support for ML and AI is highly limited on SQL and only a few companies provide that.

SQL

Picture from https://www.educba.com/sql-vs-hadoop/

Comparison (Hadoop vs. SQL)

- Learning Curve

Last but not the least, learning curve of Hadoop for entry level professionals as well as seasoned professional is moderately hard.

Hadoop

Starting with SQL systems is much easier for even entry level professionals.

SQL

Picture from https://www.educba.com/sql-vs-hadoop/

Which Database is suitable for…

Recording exam grades for student in UCONN CSE Dept?

Which Database is suitable for...

Online AI/machine learning of all car trajectories? (Uber, Lyft)

Which Database is suitable for...

Record and analyze e-commerce customer behavior offline? (Amazon, Target etc.)

Which Database is suitable for...

Realtime gaming platforms? (WOW, PUBG, LOL, Dota2)

Which Database is suitable for…

Messaging system (IMessager, Wechat)