INTRODUCTION TO BIG DATA ANALYTICS

Quách Đình Hoàng

Outline

- A historical review for Big Data
- 3Vs-6Vs characteristics of Big Data
- Machine Learning (ML)
- Big Data and cloud computing
- Hadoop, Hadoop distributed file system (HDFS), MapReduce, Spark
- BDA = ML + CC (Cloud Computing)

References

- Caesar Wu, Rajkumar Buyya, and Kotagiri Ramamohanarao, *Big Data Analytics = Machine Learning + Cloud Computing*, In *Big Data: Principles and Paradigms*, Morgan Kaufmann, 2016. http://www.cloudbus.org/papers/BigDataAnalytics2016.pdf

A Short History of Big Data (1)

1997, The problem of Big Data, NASA researchers, Michael Cox et and David Ellsworth's paper

1998, Google was founded

1999, Apache Software Foundation (ASF) was established

2000, Doug Cutting launched his indexing search project: Lucene

2000, L Page and S. Brin wrote paper "the Anatomy of a Large-Scale Hyertextual Web search engine"

2001, The 3Vs, Doug Laney's paper "3D data management: controlling data Volume, Velocity & Variety"

2002, Doug Cutting and Mike Caffarella started Nutch, a subproject of Lucene for crawling websites

2003, Sanjay Ghemawat et al. published "The Google File System"(GFS)

2003, Cutting and Caffarella adopted GFS idea and create Nutch Distribute File System (NDFS) later, it became HDFS

2004, Google Began to develop Big Table

2004, Yonik Seeley created **Solr** for Text-centric, read-dominant, document-oriented & flexible schema search engine

2004, Jeffrey Dean and Sanjay Ghemawat published "Simplified Data Processing on Large Cluster" or MapReduce

2005 Nutch established Nutch MapReduce

2005, Damien Katz created Apache CouchDB (Cluster Of Unreliable Commodity Hardware), former Lotus Notes

A Short History of Big Data (2)

2006, Cutting and Cafarella started Hadoop or a subproject of Nutch

2006, Yahoo Research developed Apache Pig run on Hadoop

2007, 10gen, a start-up company worked on Platform as a Service (PaaS). Later, it became MongoDB

2007, Taste project

2008, Apache Hive (extend SQL), HBase (Manage data) and Cassandra(Schema free) to support Hadoop

2008, Mahout, a subproject of Lucene integrated Taste

2008 Hadoop became top level ASF project

2008 TUB and HPI initiated Stratosphere Project and later become Apache Flink

2009, Hadoop combines of HDFS and MapReduce. Sorting one TB 62 secs over 1,460 nodes

2010, Google licenced to ASF Hadoop

2010, Apache Spark , a cluster computing platform extends from MapReduce for in-memory primitives

2011, Apache Storm was launched for a distributed computation framework for data stream

2012, Apache Dill for Schema-Free SQL Query Engine for Hadoop, NoSQL and cloud Storage

2012, Phase 3 of Hadoop – Emergence of "Yet Another Resource Negotiator"(YARN) or Hadoop 2

2013 Mesos became a top level Apache project

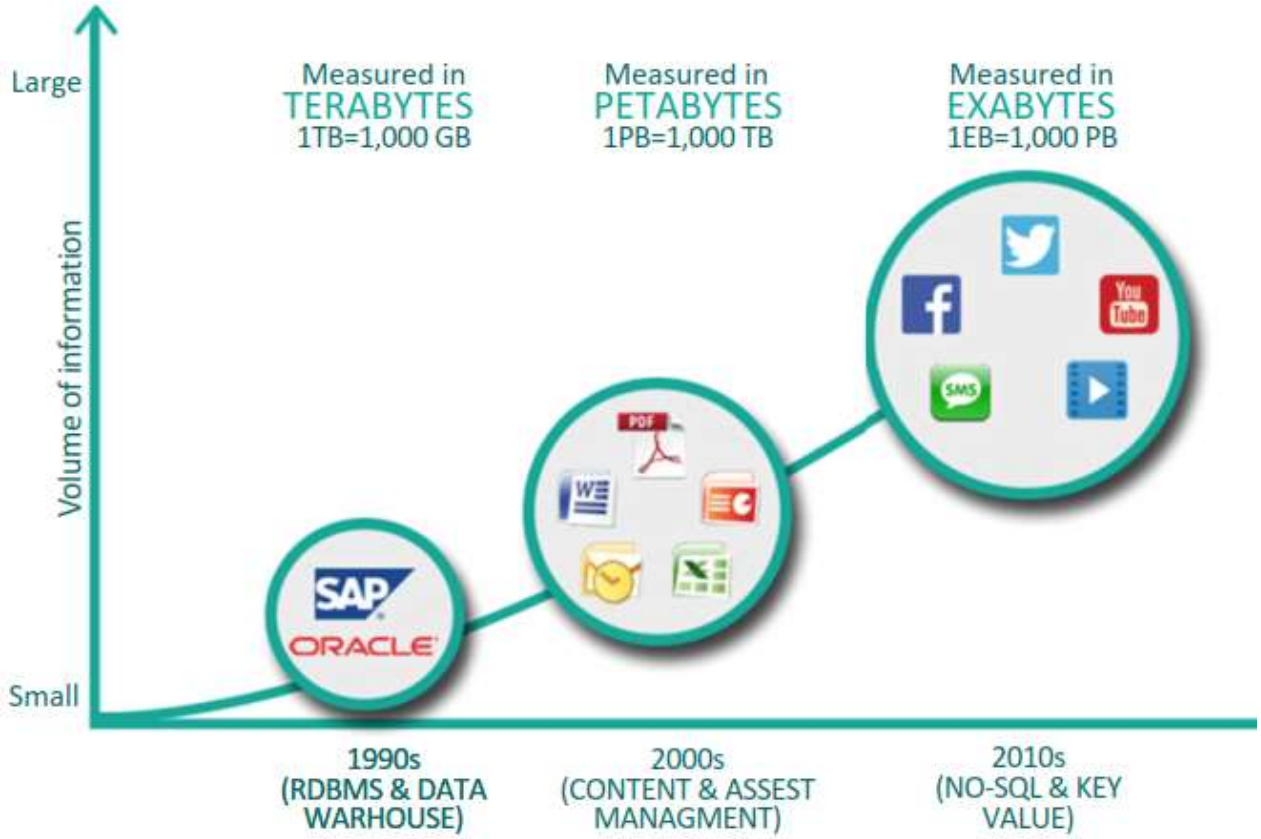
2014, Spark has > 465 contributors in 2014, the most active ASF project

2015, Enter Zeta Byte Era

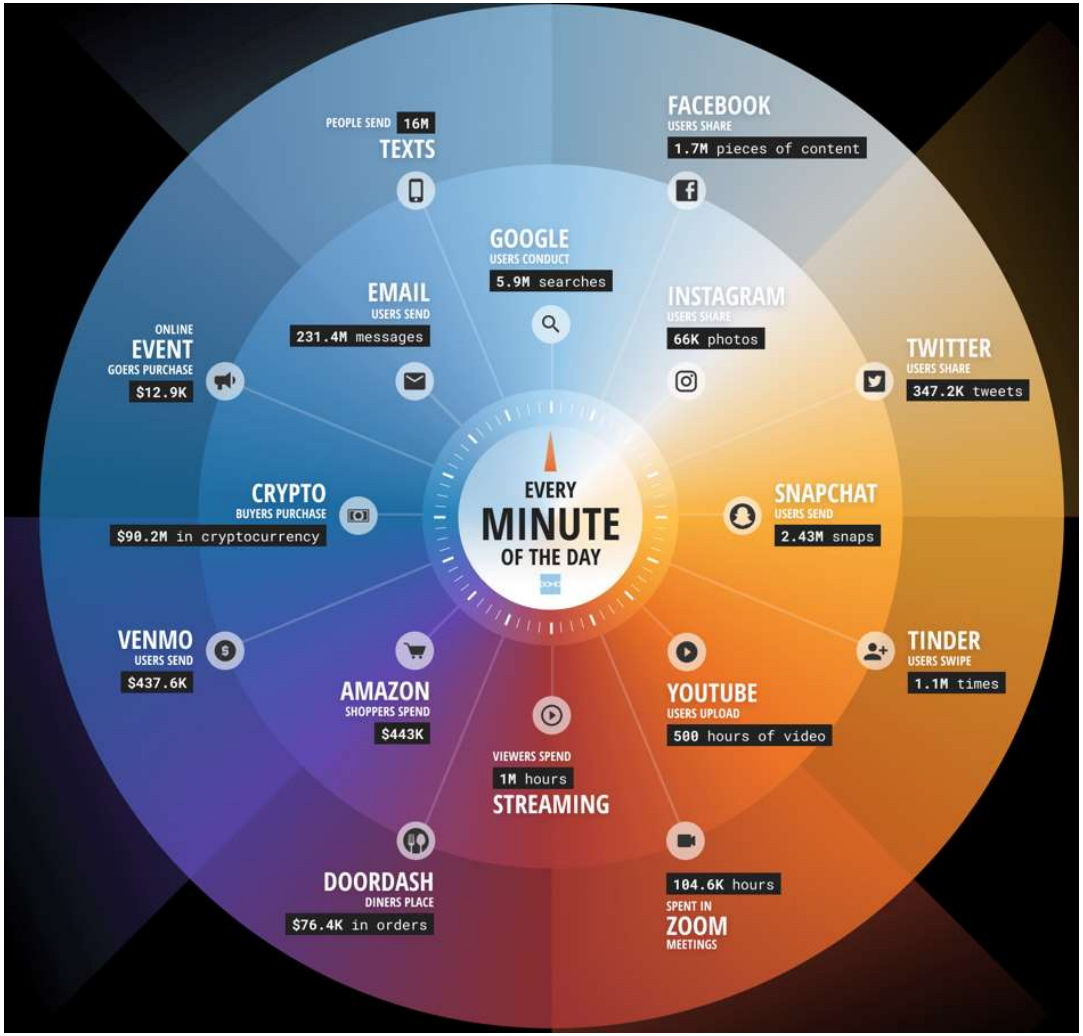
Typical Size of Different Data Files

Media	**Average Size**	**Notes (2014)**
Web page	1.6–2 MB	Average 100 objects
eBook	1–5 MB	200–350 pages
Song	3.5–5.8 MB	Average 1.9 MB/per minute (MP3) 256 Kbps rate (3 mins)
Movie	100–120 GB	60 frames per second (MPEG-4 format, Full High Definition, 2 hours)

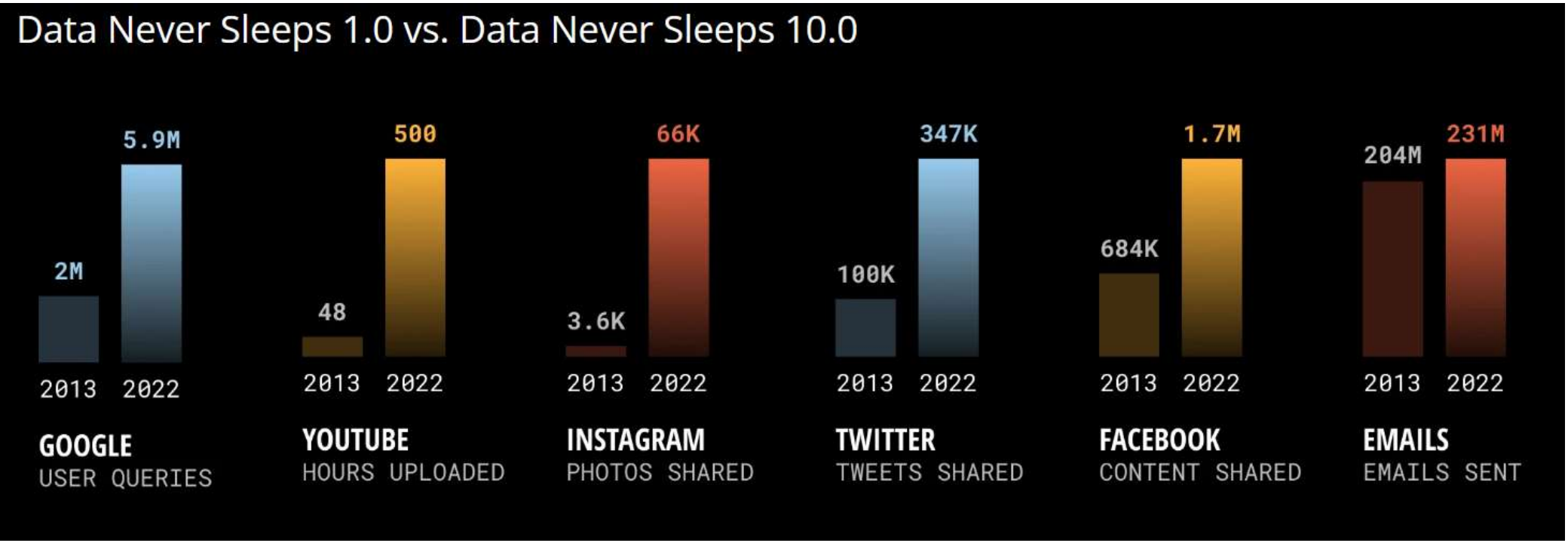
The data evolution over the years



Big Data Phenomenon - Data Never Sleep

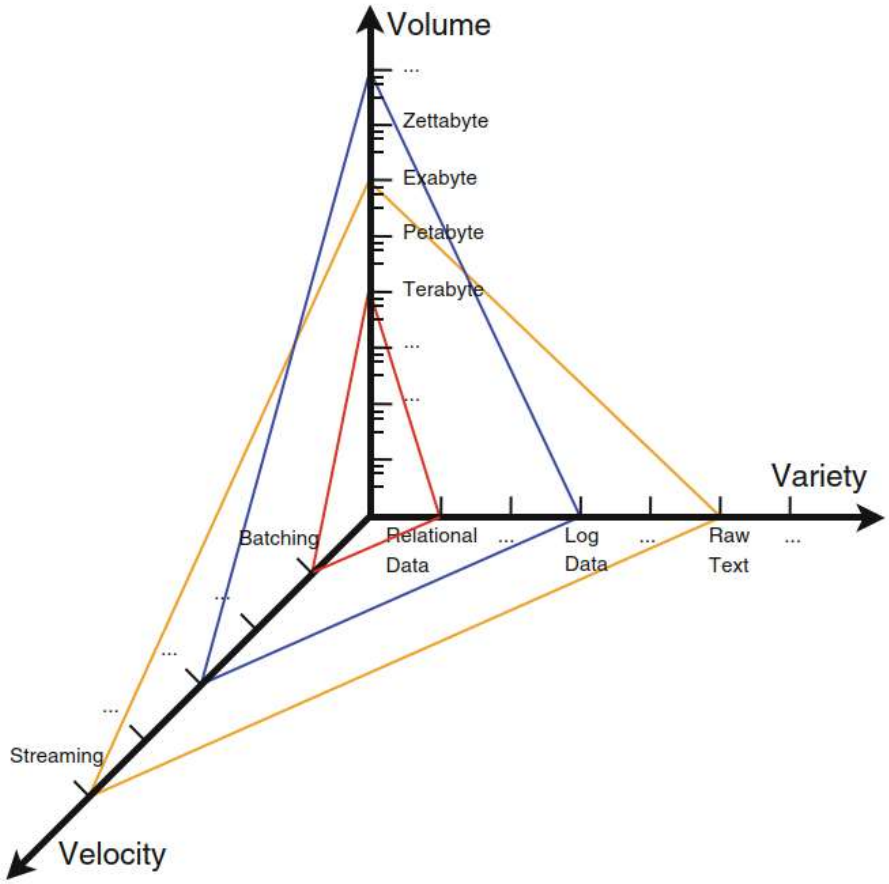


Source: https://www.domo.com/data-never-sleeps

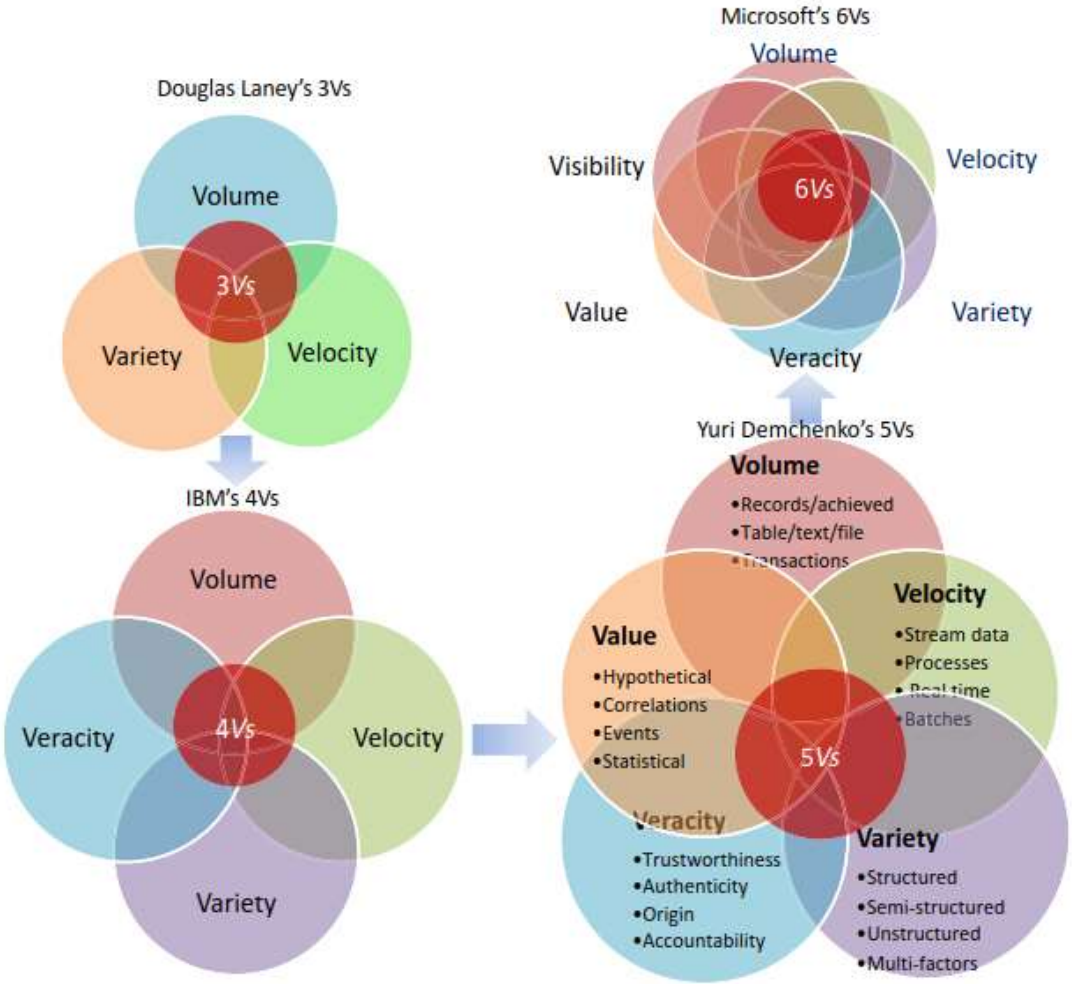


Source: https://www.domo.com/data-never-sleeps

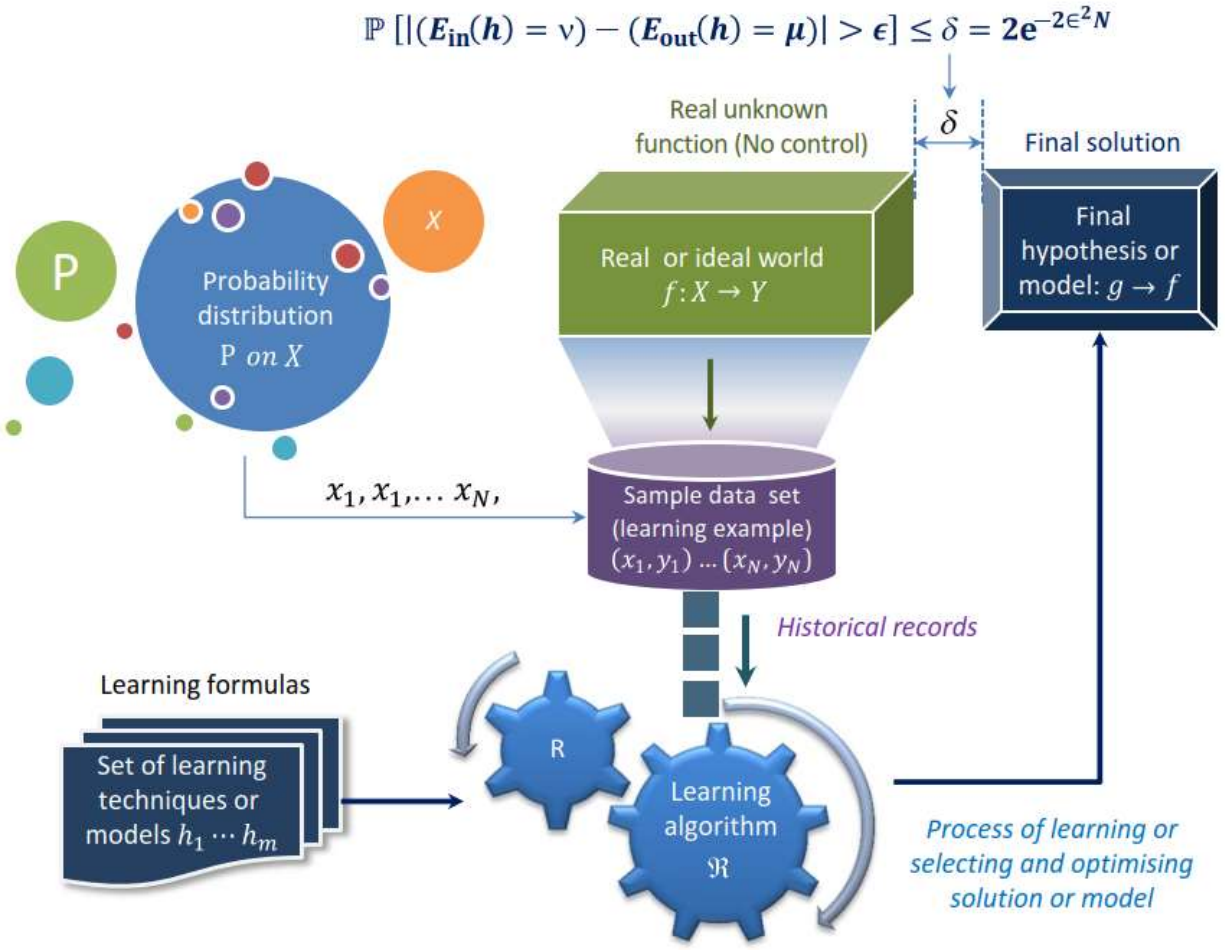
3V Characteristics of Big Data



3-6Vs Characteristics of Big Data

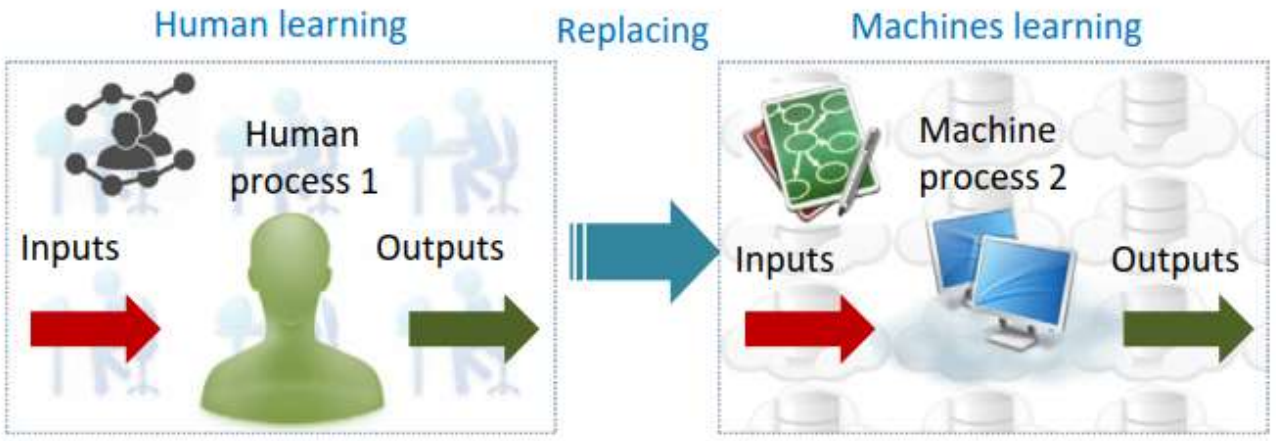


Machine learning process



Replacing humans in the learning process

- The ultimate goal of ML is to build systems that are of at the level of human competence in performing complex tasks

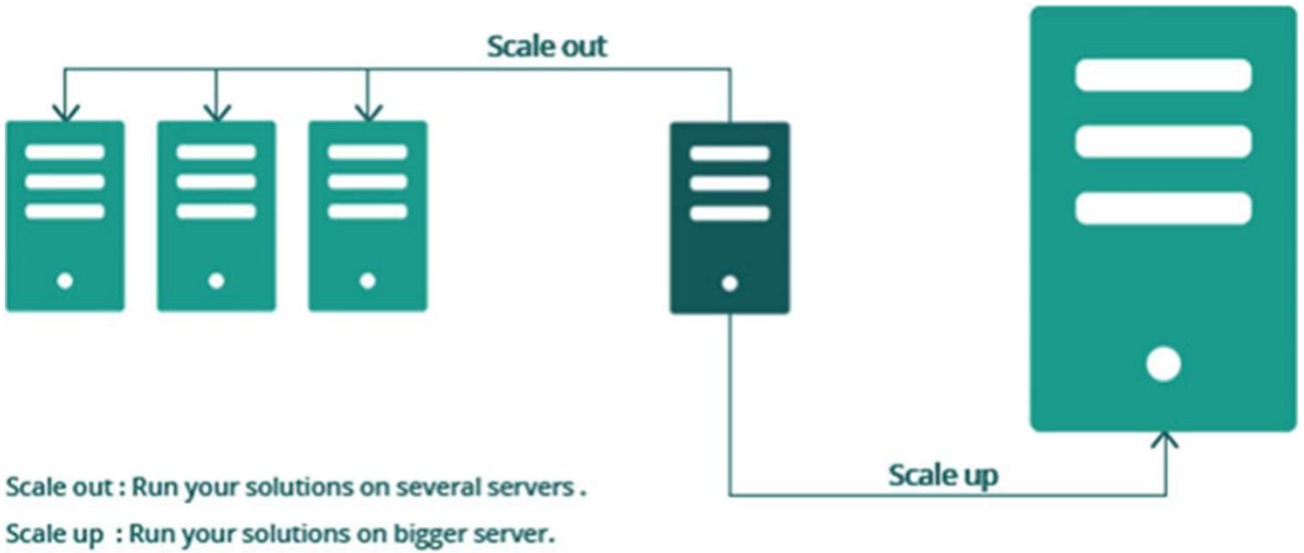


Big Data Analytics and Cloud Computing

- Cloud Computing (CC) plays a critical role in the Big Data Analytics (BDA) process
  - it offers subscription-oriented access to computing infrastructure, data, and application services
- The original objective of BDA was to leverage commodity hardware to build computing clusters and scale-out the computing capacity
  - Cost: enable many small to medium companies to implement BDA (pay as you go)
  - Scalability: almost “infinite” capacity
  - Elasticity: easily scale-out and scale down

Scale out vs. scale up

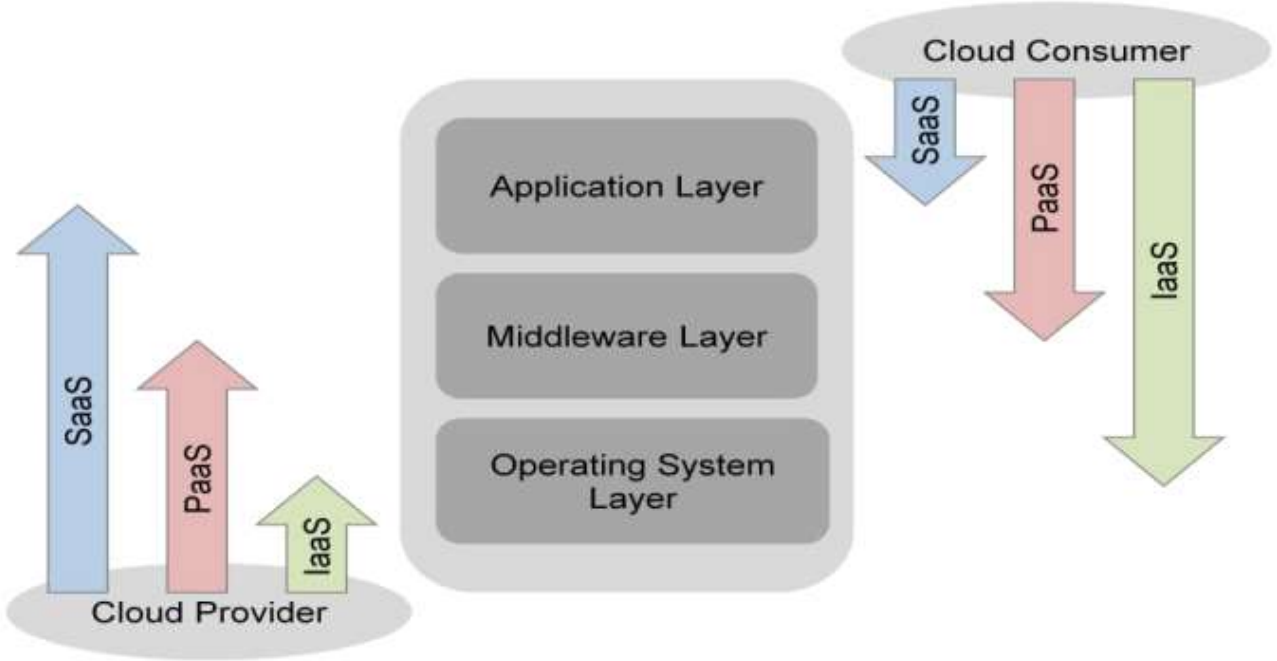
- Scale out = horizontal scale
- scale up = vertical scale



Cloud computing services

- Infrastructure as a Service (IaaS)
  - Serve computing resources: CPU, storage, networks, …
  - Amazon EC2, Rackspace, …
- Platform as a Service (PaaS)
  - Serve API, maintenance, upgrades
  - Google App Engine, Apple Play Store, …
- Software as a Service (SaaS)
  - Serve applications
  - Gmail, Dropbox, …

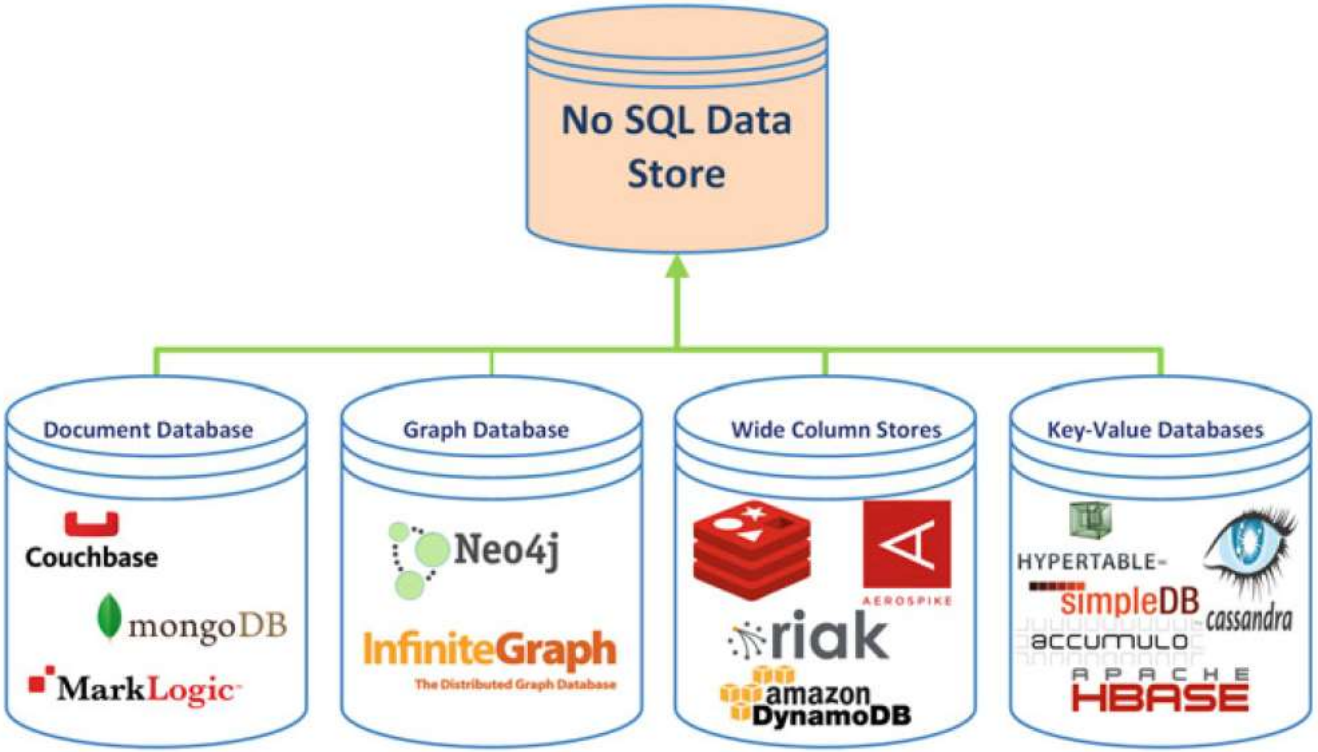
Scope of Controls between Provider and Consumer



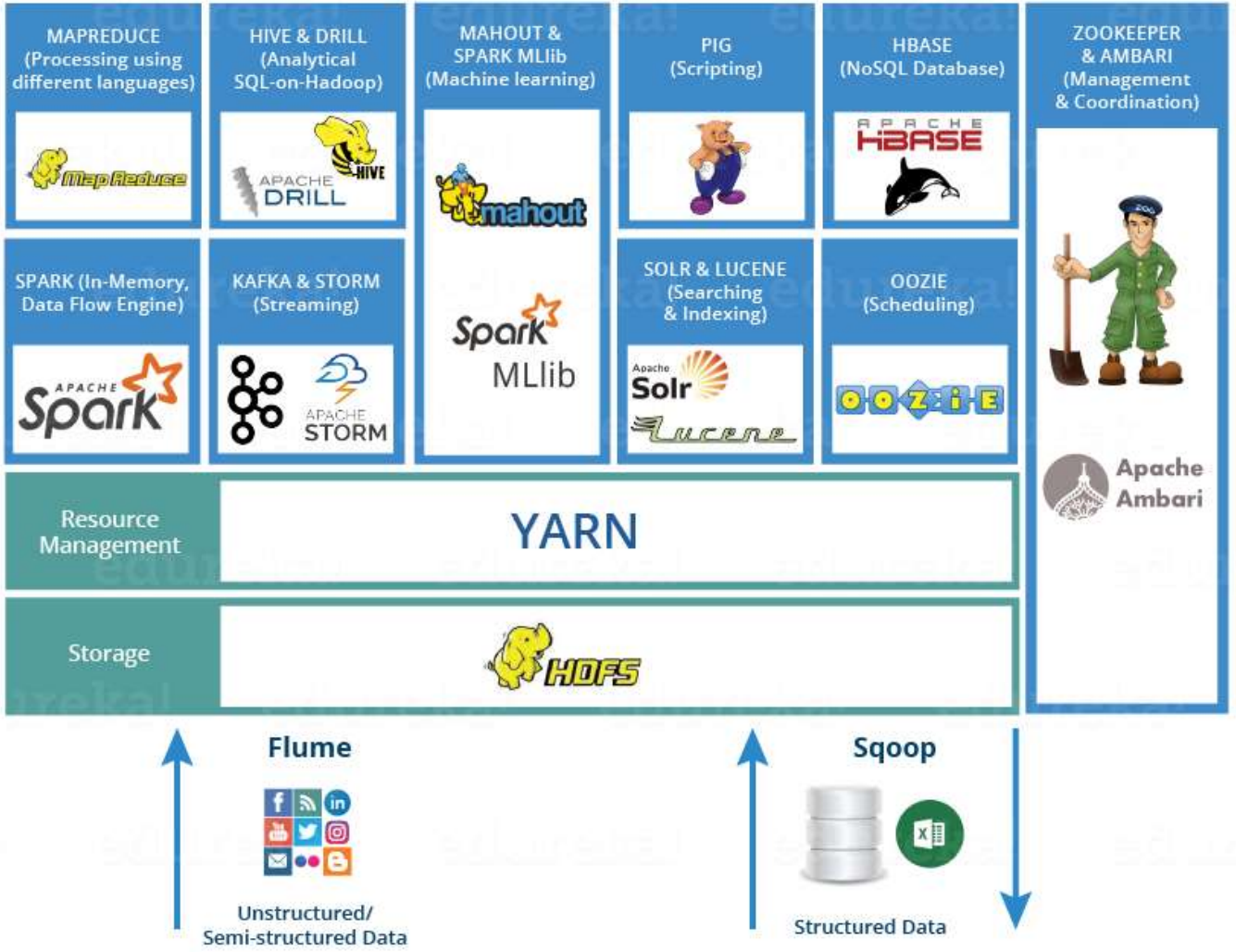
Big Data Storage Systems

- Structured data: Data with a defined format and structure
  - CSV files, spreadsheets, traditional relational databases, and OLAP data cubes
- Semi-structured data: Textual data files with a flexible structure that can be parsed
  - XML, JSON
- Unstructured data: Data that have no inherent structure
  - text documents, images, PDF files, and videos

Types of NoSQL data stores



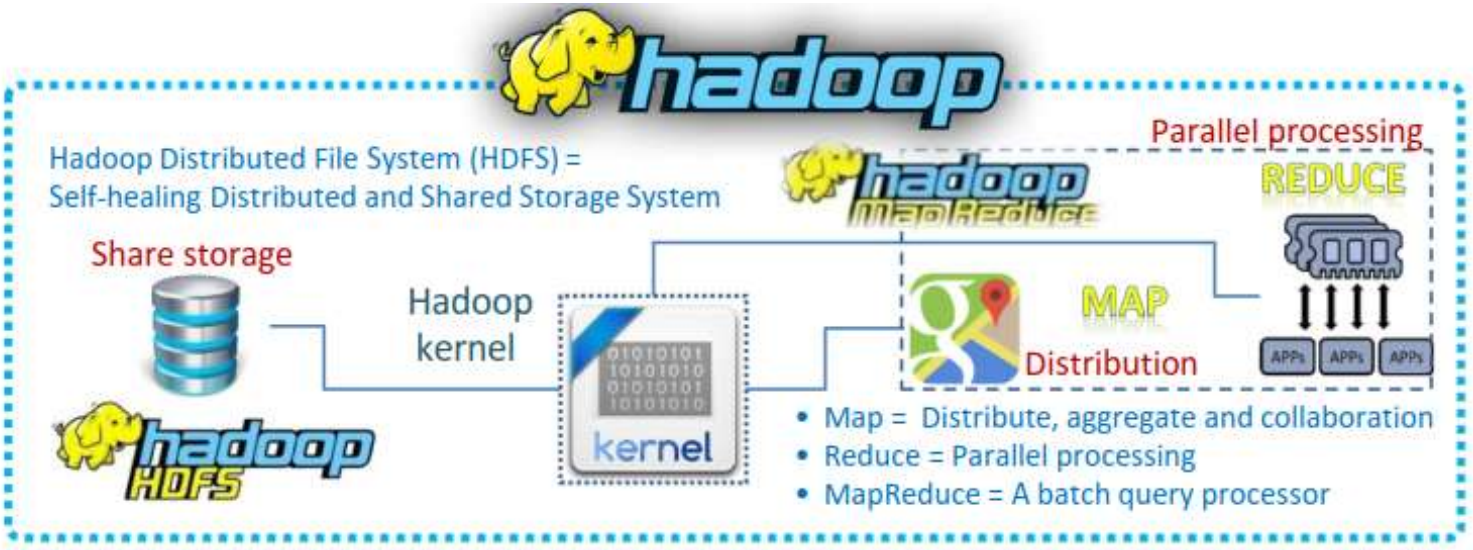
Hadoop ecosystem



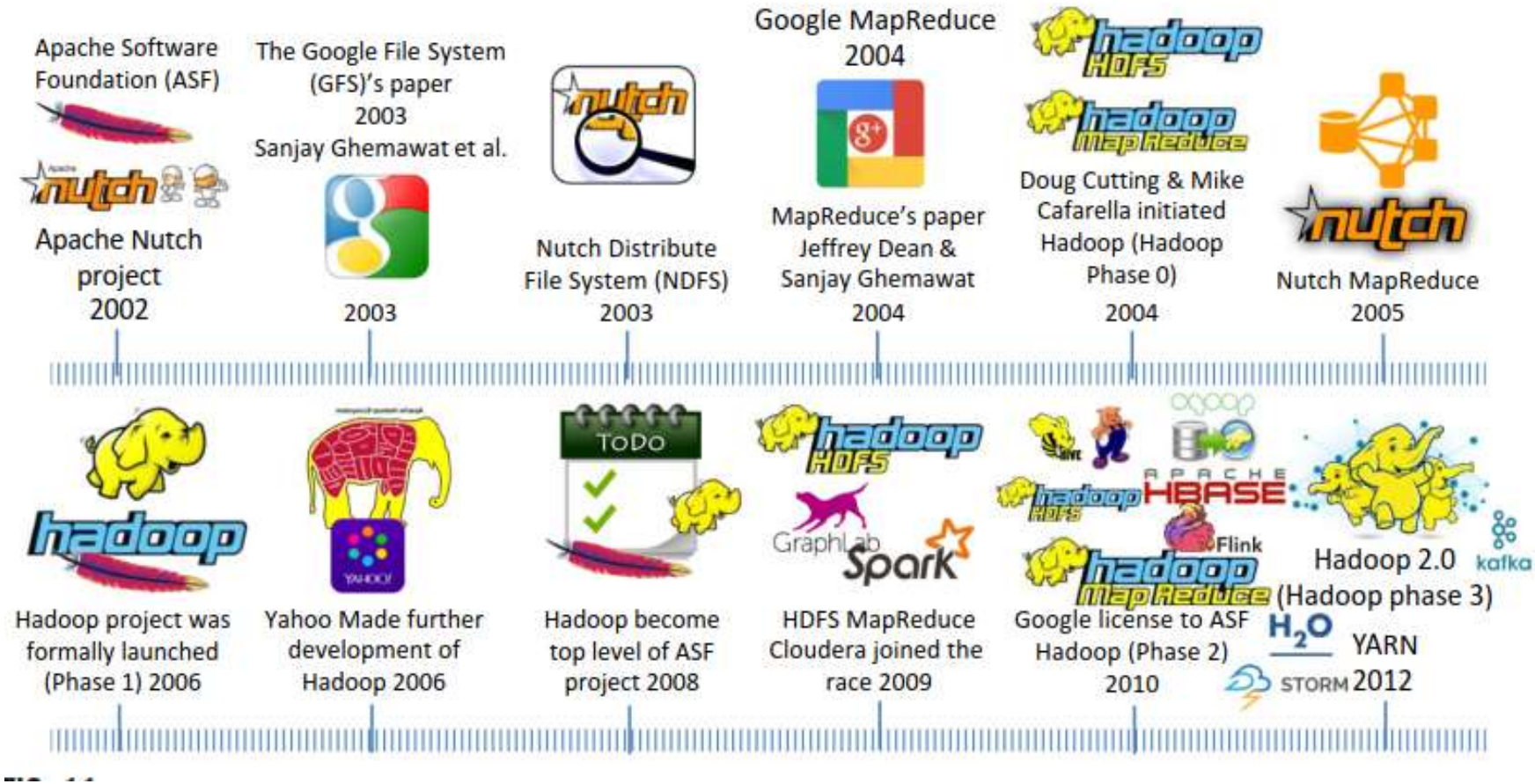
Hadoop Technology Stack and Ecosystem

Hadoop kernel

- HDFS (file storage), Map (distribute function), and Reduce (parallel processing function)

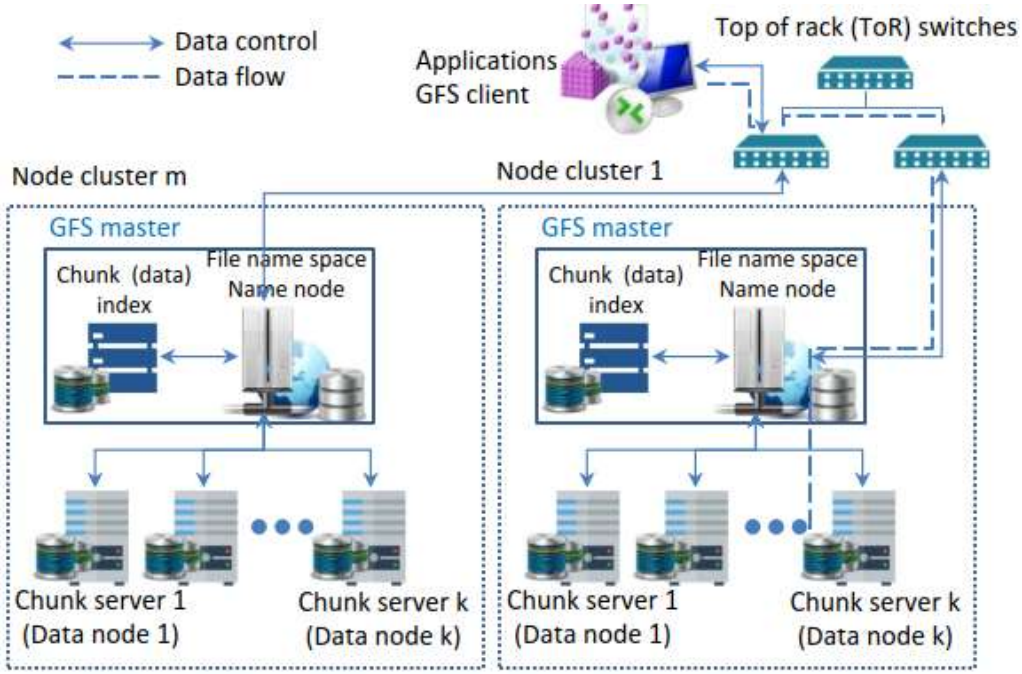


Briefing history of Hadoop

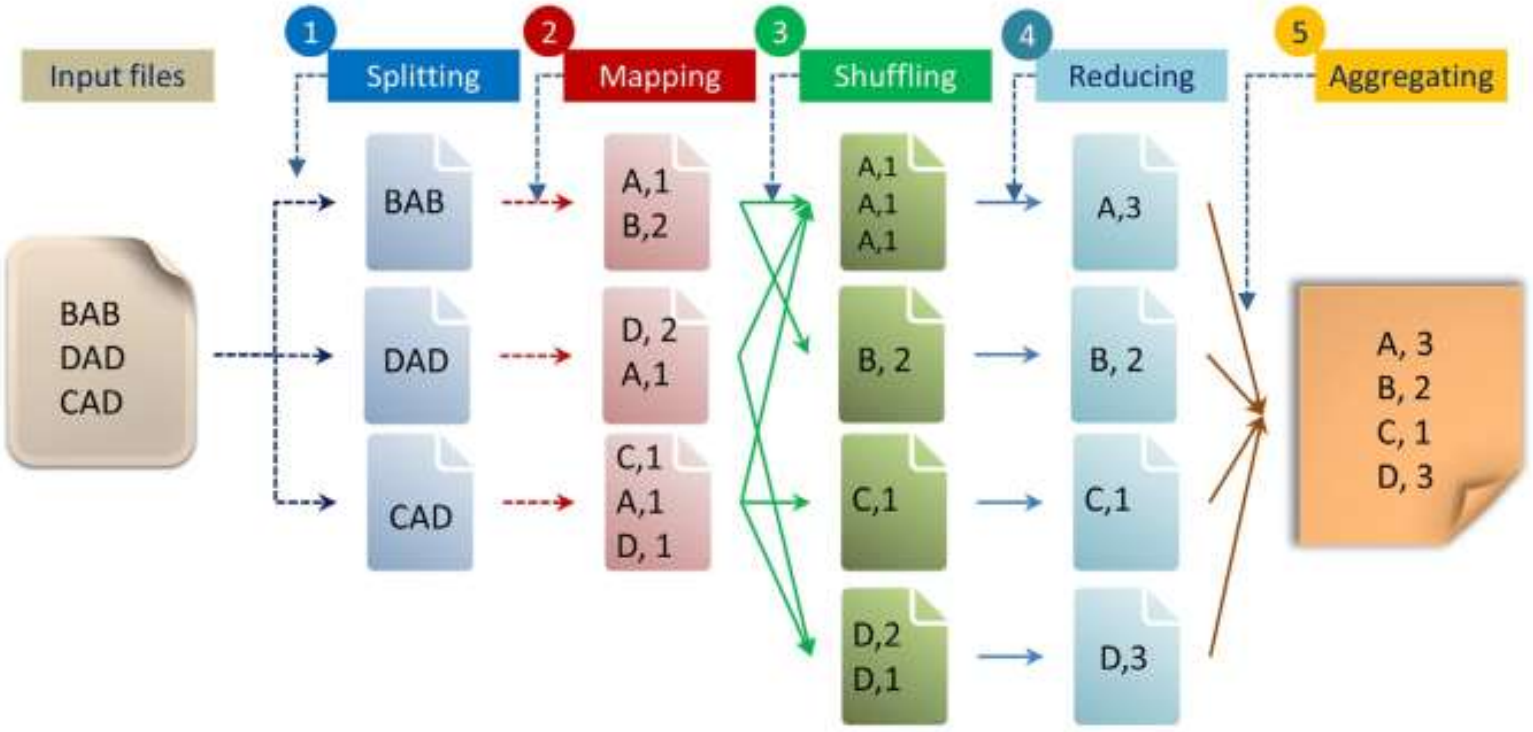


Google file system (GFS)

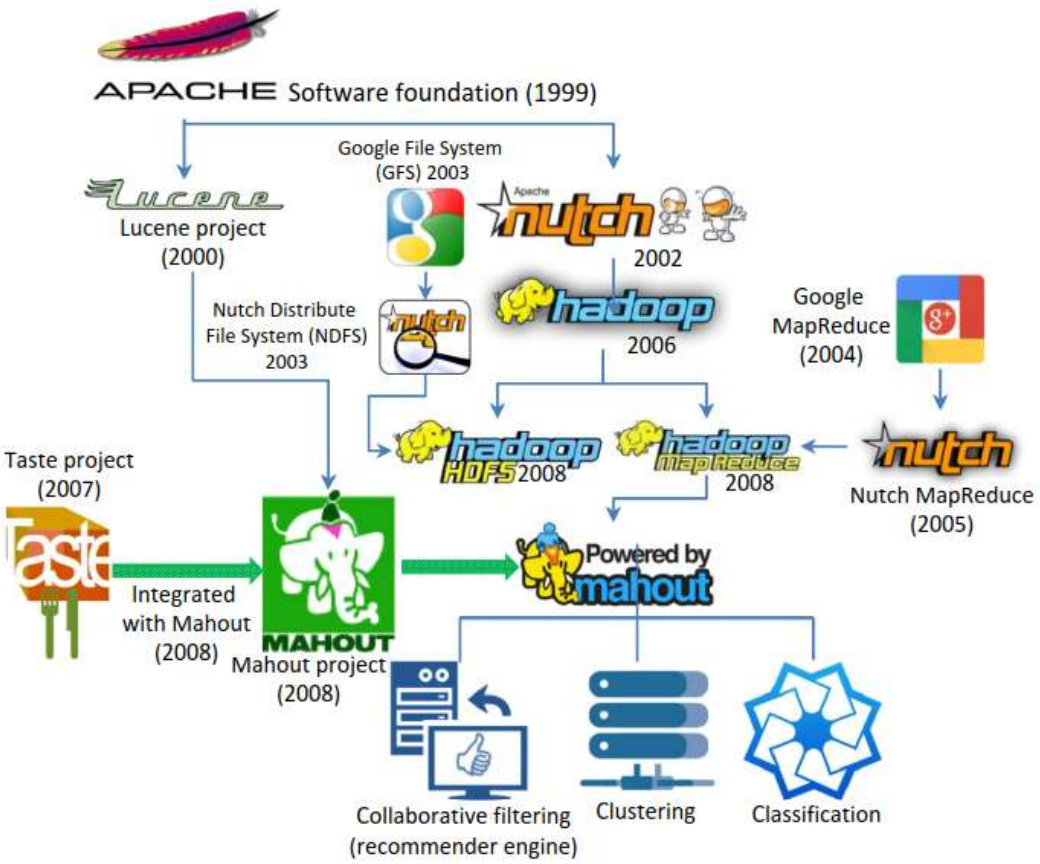
- The GFS architecture consists of three components
  - Single master server (or name node for Hadoop)
  - Multiple chunk servers (or data nodes for Hadoop)
  - Multiple clients



MapReduce programming model



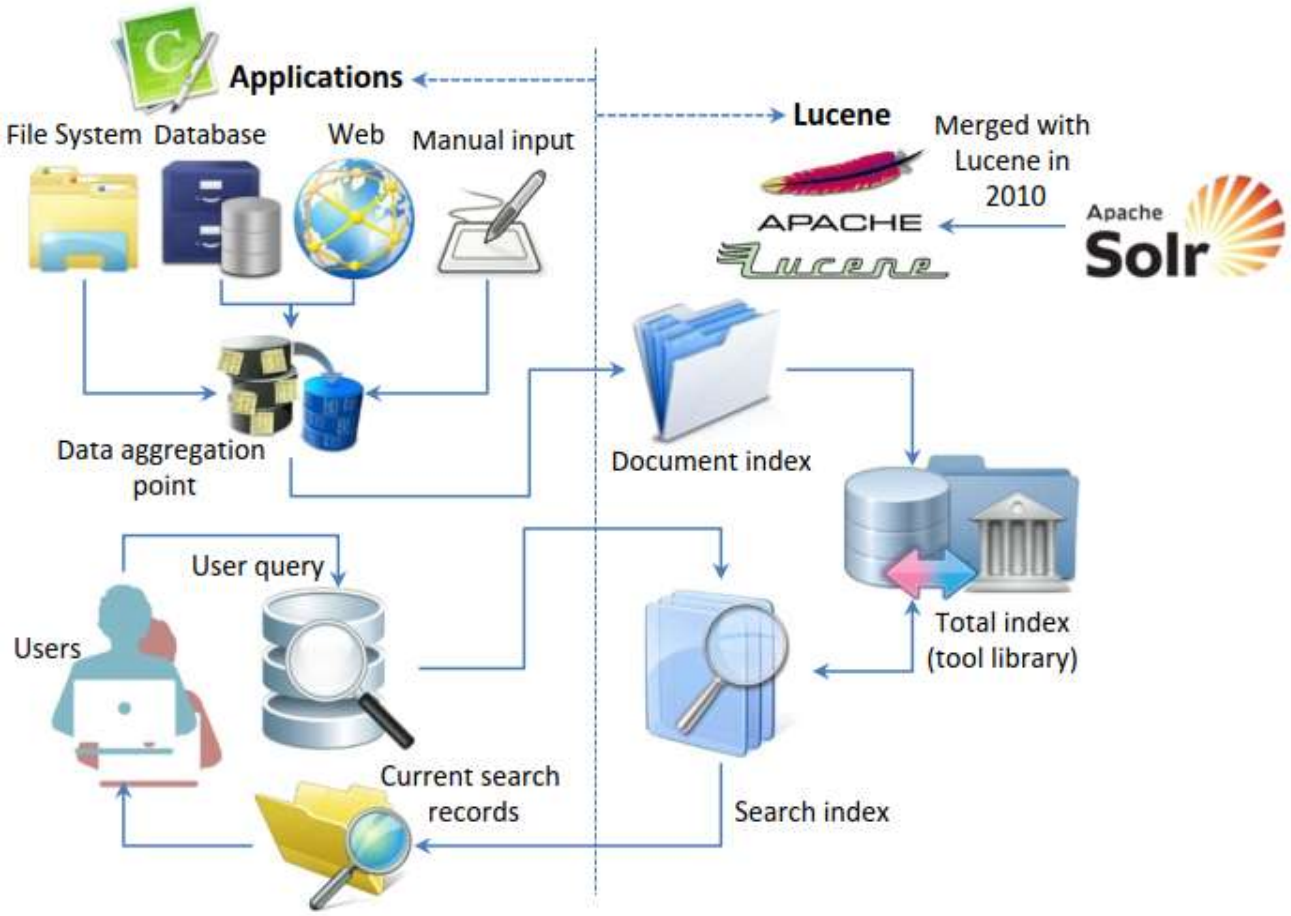
Evolution of GFS, HDFS MapReduce, and Hadoop



The origin of Hadoop project

- Lucene
  - a high-performance scalable information retrieval (IR) library
  - was written by Doug Cutting in 2000 in Java
  - In Sep. 2001, Lucene was absorbed by ASF
- Nutch
  - Nutch is the predecessor of Hadoop, built by Doug Cutting in 2002
  - There are two main reasons to develop Nutch
    - Create a Lucene index (web crawler)
    - Assist developers to make queries of their index
- Mahout
  - a Java-based ML library that covers all ML algorithms
    - Collaborative filtering (recommender engines)
    - Clustering
    - Classification

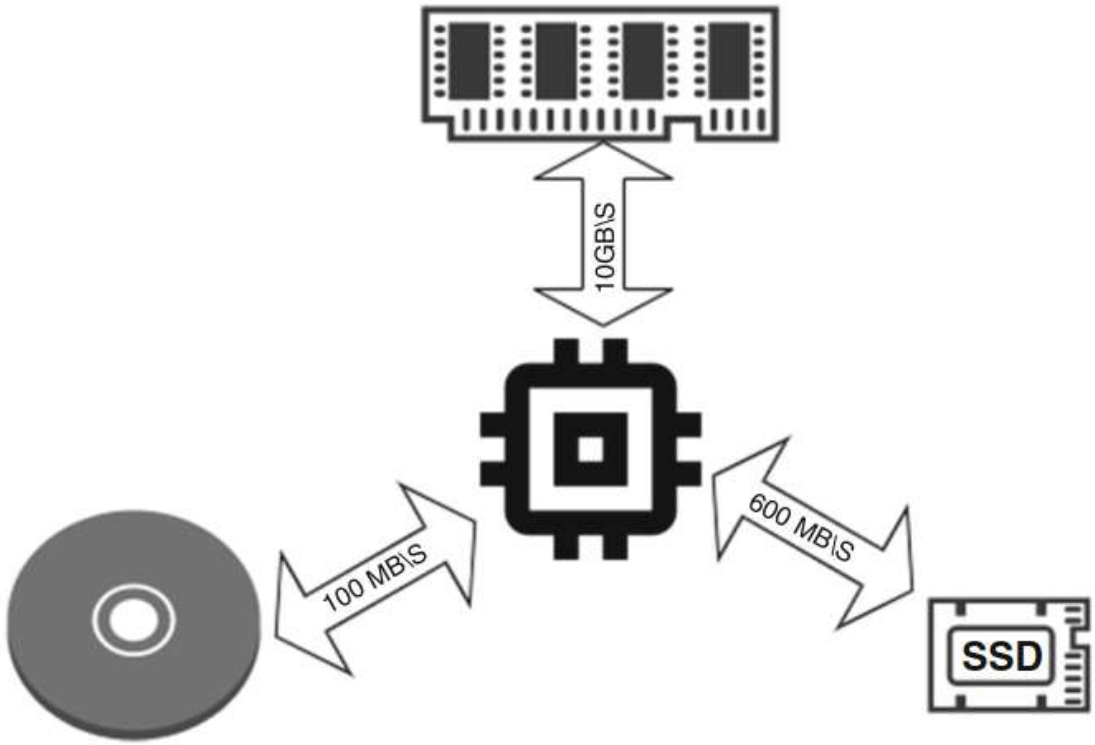
Apache Lucene



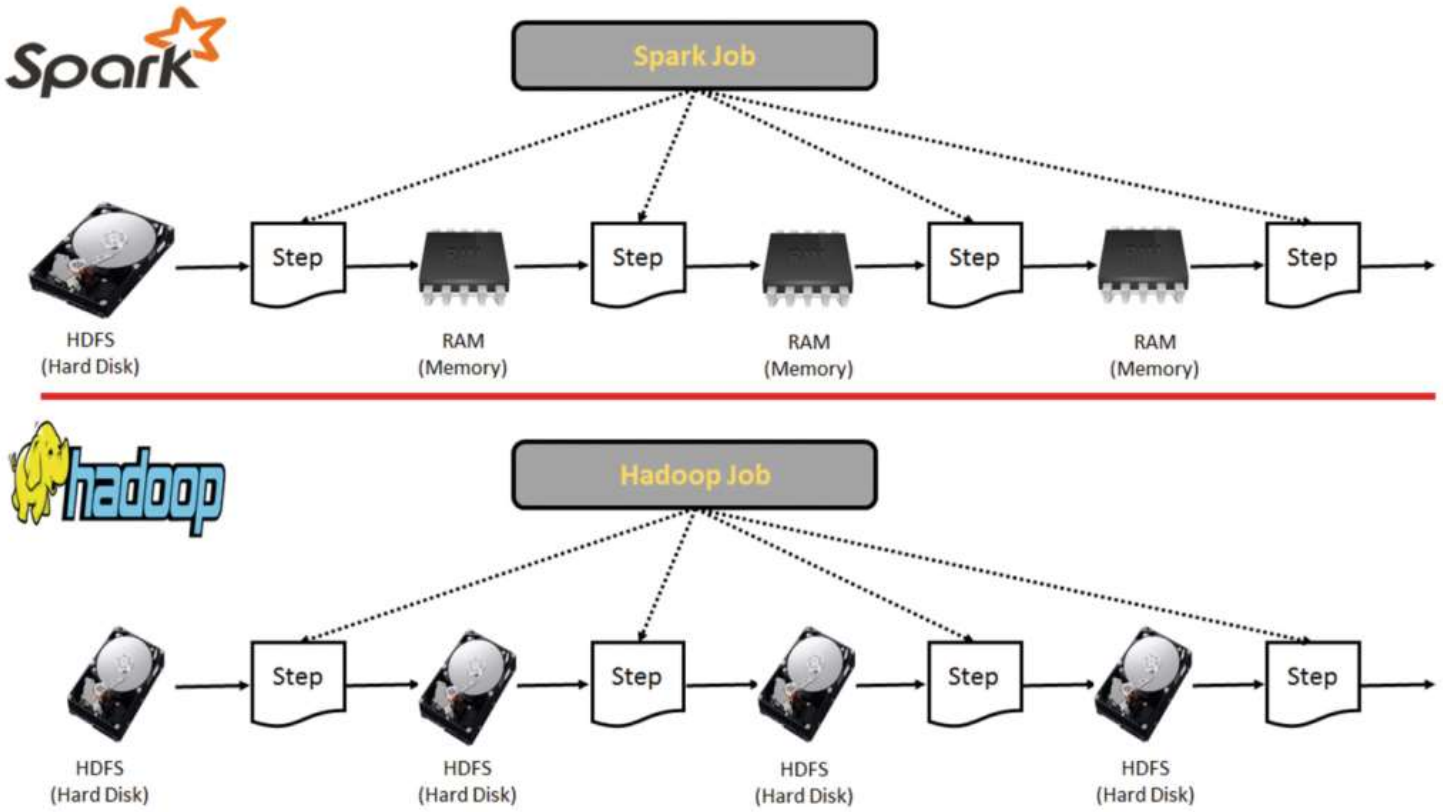
Spark

- Spark was developed by the UC Berkeley AMP Lab
- The main contributor is Matei Zaharia et al.
- It intends to replace MapReduce model with a better solution
- It would be 10-20 times faster than MapReduce for certain type of workload
- Although it attempts to replace MapReduce, it leverages Hadoop's file storage system

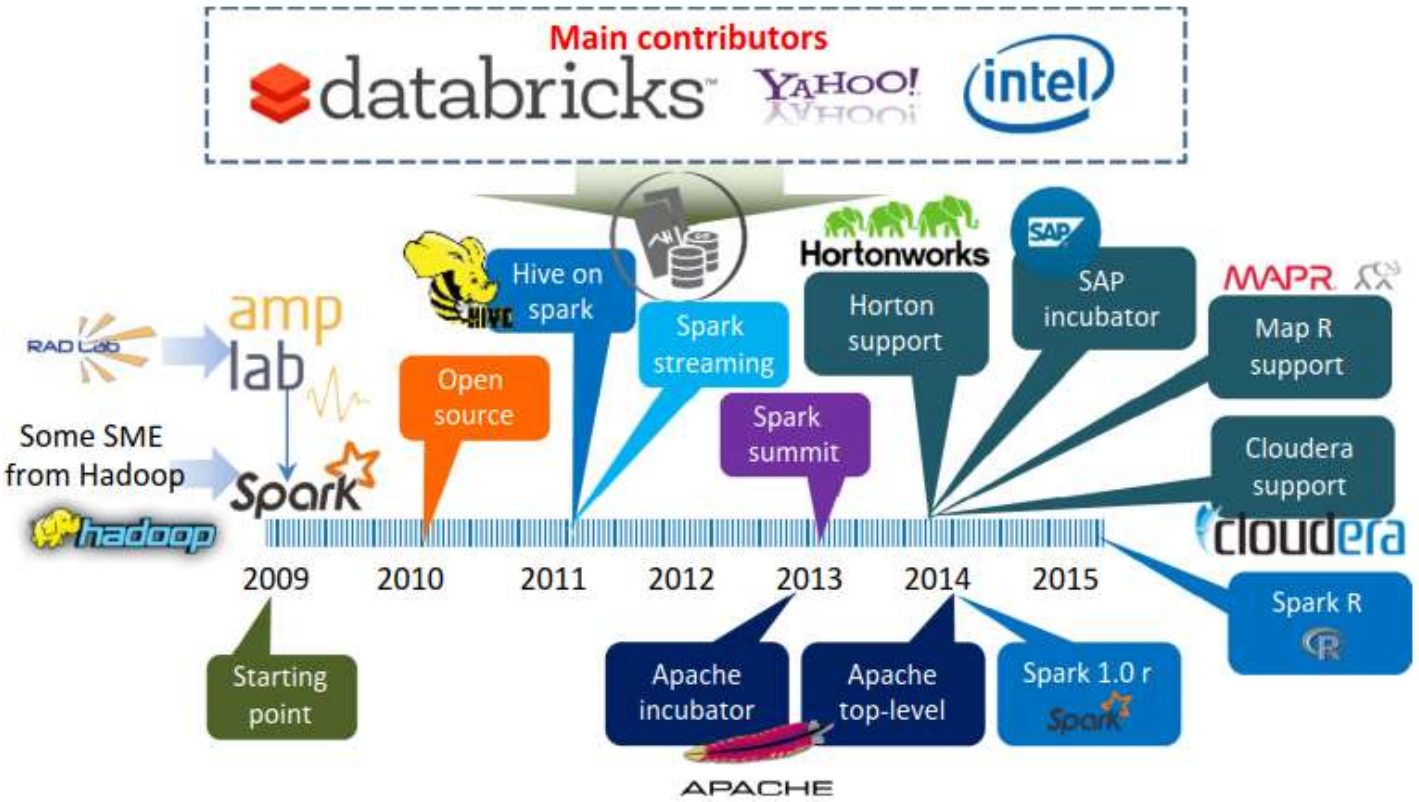
Differences on data transfer speed



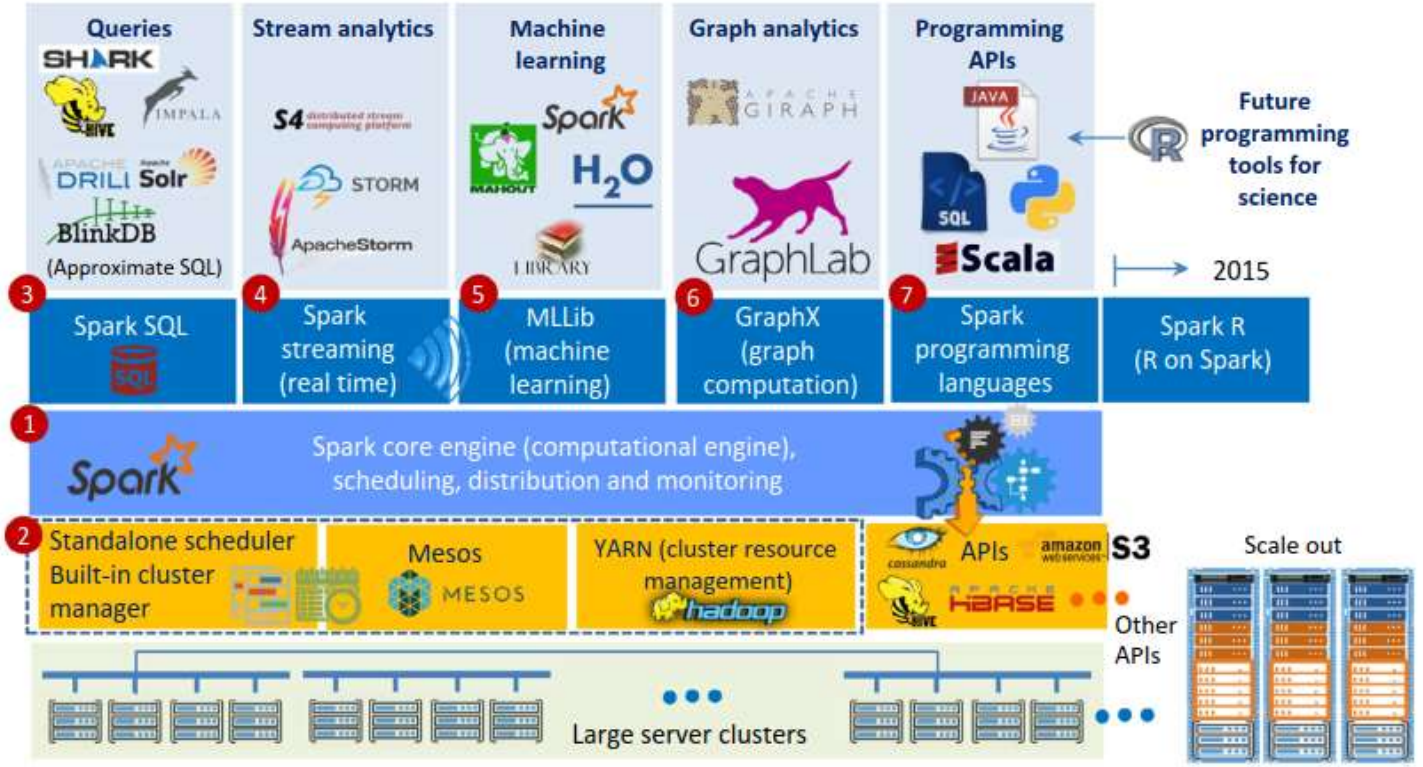
Spark framework vs Hadoop framework



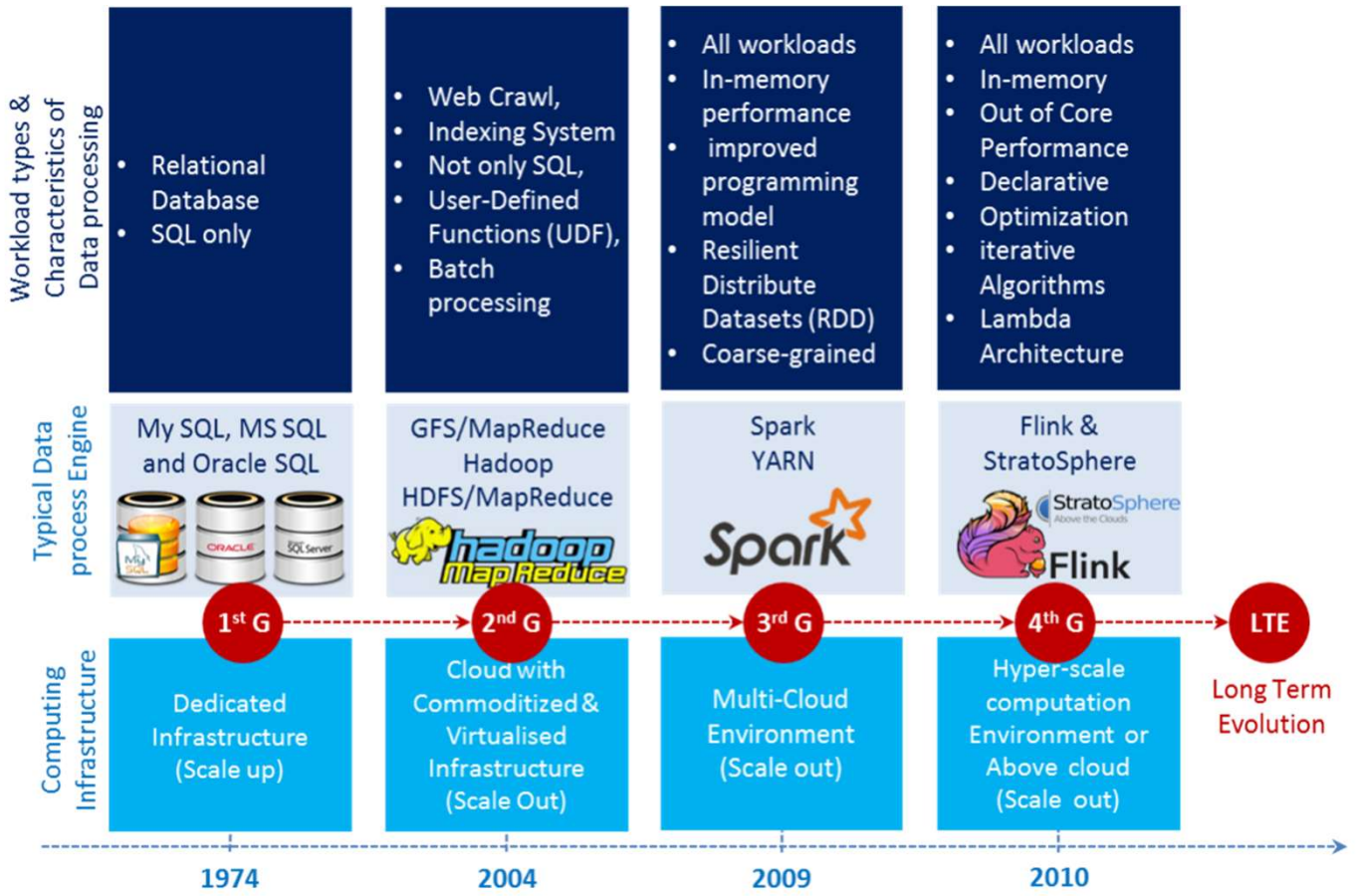
Spark history



Spark analytic stack



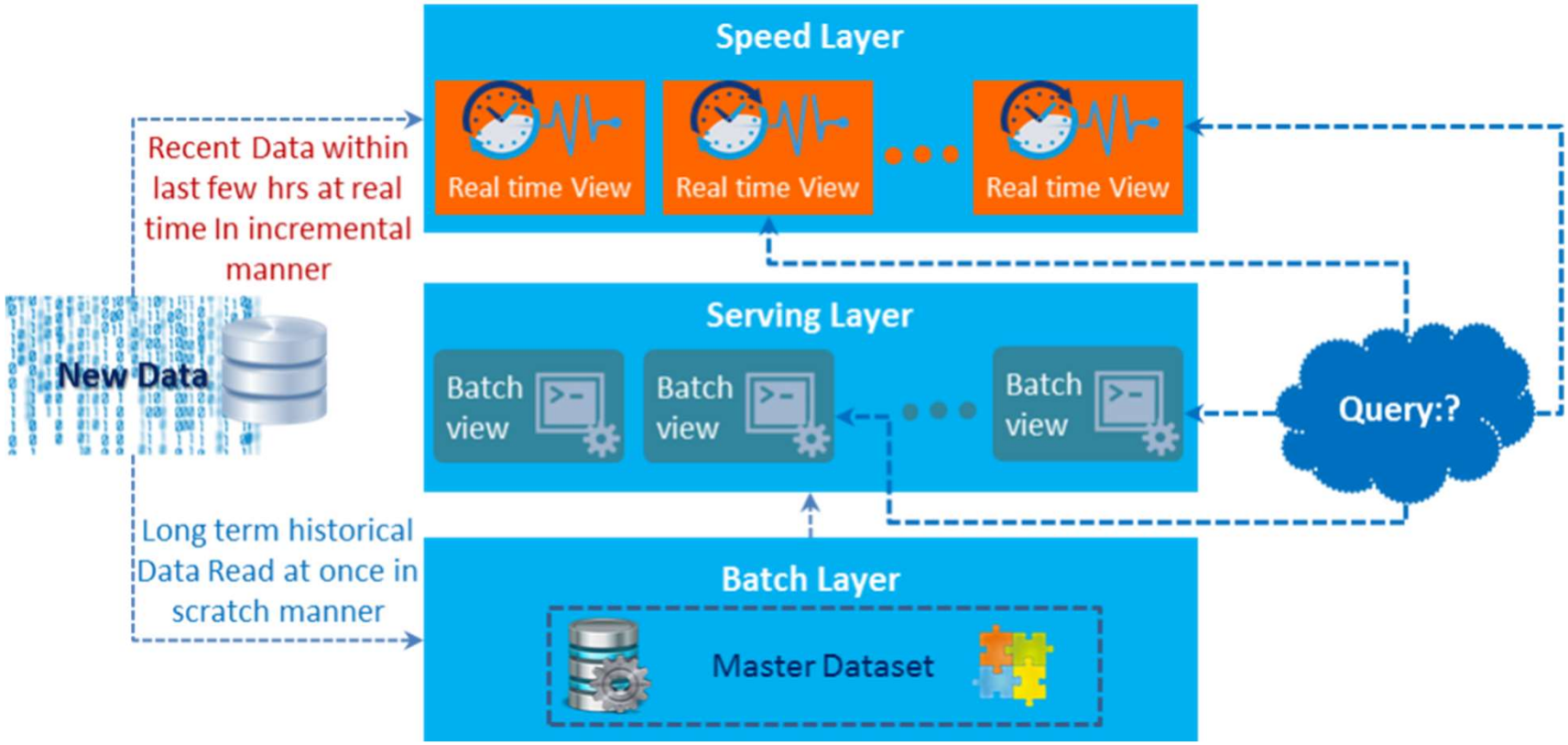
Evolution of Data and Big Data process engines



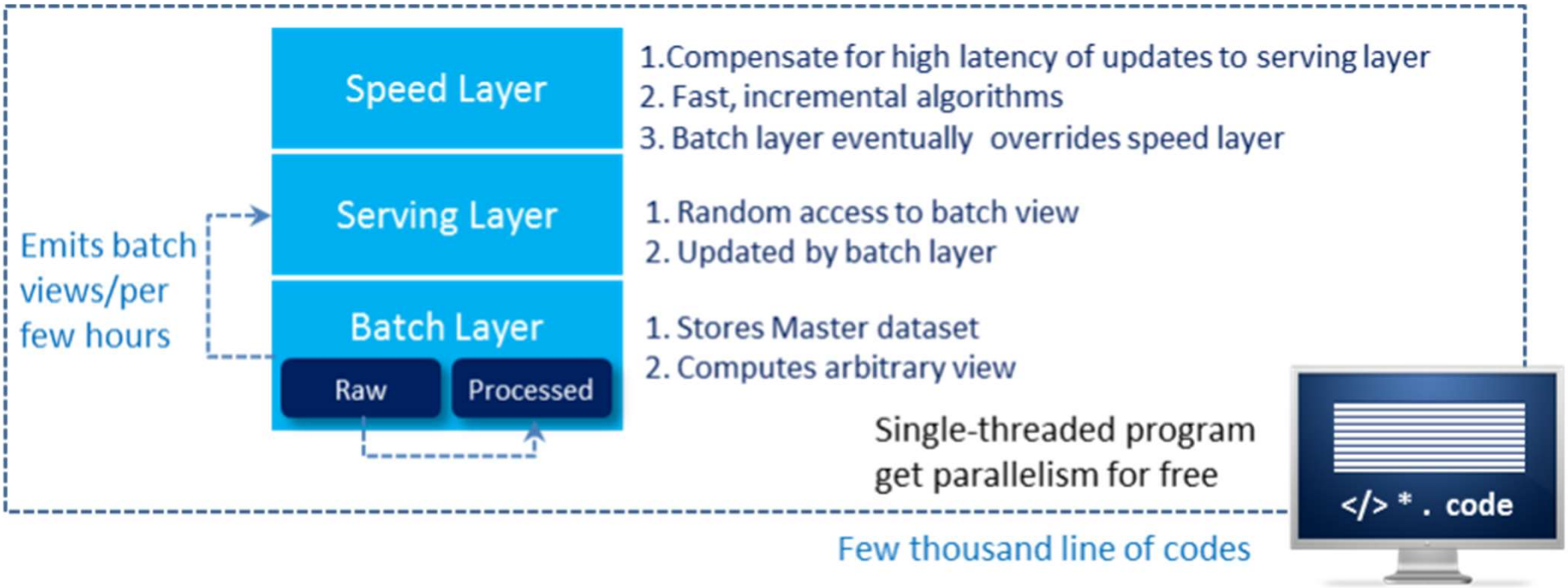
Data Processing Engine Comparison

Data process engines comparison	MapReduce	Tez	Spark	Flink
Start at	2004	2007	2009	2010
API	MapReduce on Key/Value pairs	Key/Value pair Readers/Writers	Transformations on key/value pair collections	Iterative transformations on collection or iteration aware
Paradigm	MapReduce	Direct Acyclic Graph (DAG)	Resilient Distributed Datasets (RDD)	Cyclic data flows or dataflow with feedback edges
Optimization	none	none	Optimization of SQL queries	Optimization in all APIs
Execution	Batch	Batch sorting and partitioning	Batch with memory pinning	Stream with out of core algorithms
Enhanced features plus Specialise particular workloads	• Small recoverable tasks, • Sequential code inside map & reduce functions	• Extends map/reduce model to DAG model • Backtracking-based recovery	• Functional implementation of Dryad recovery (RDDs) • Restrict to coarse-grained transformations • Direct execution of API	• Embed query processing runtime in DAG engine • Extend DAG model to cyclic graphs • Incremental construction of graphs

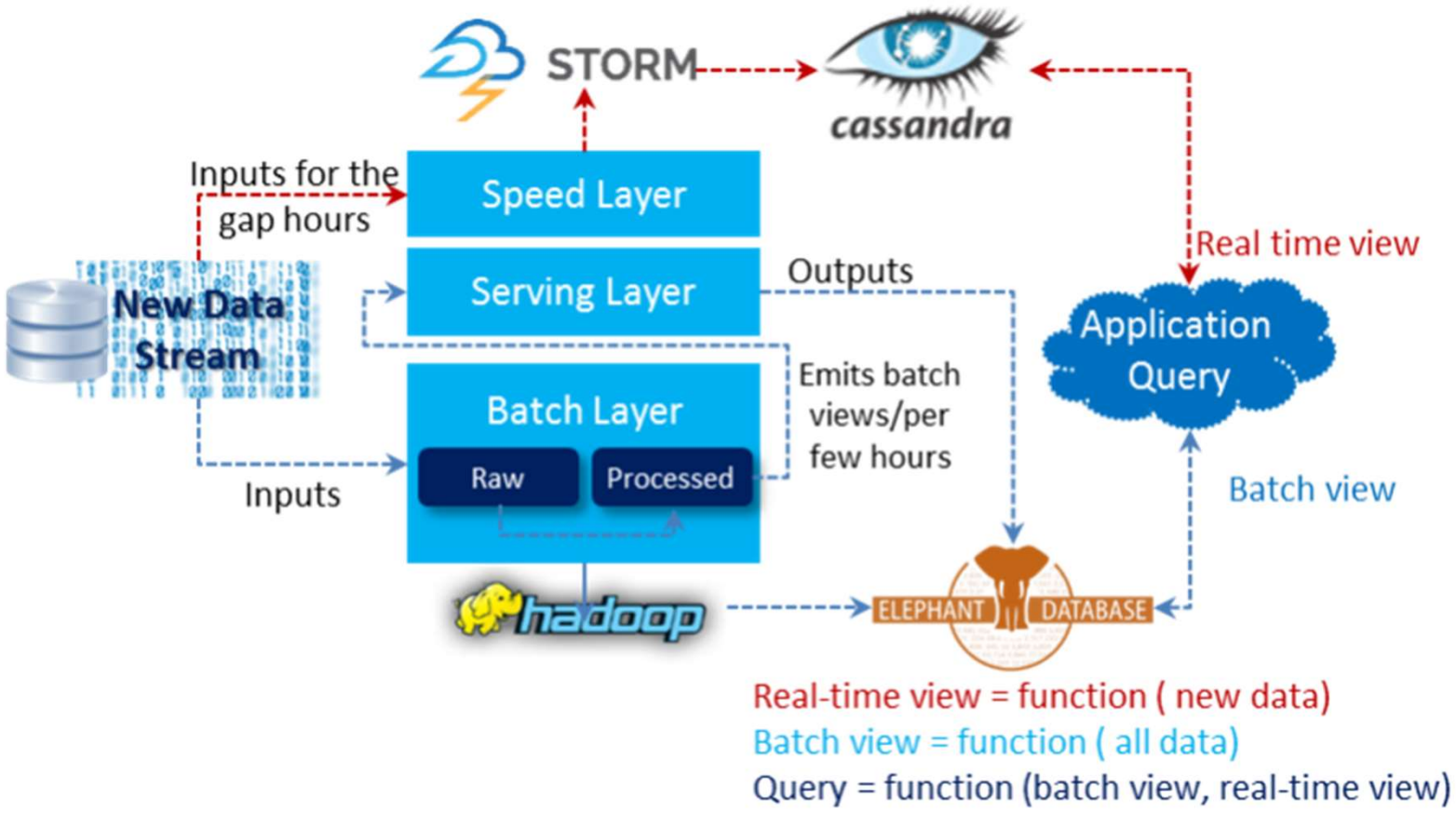
Lambda Architecture



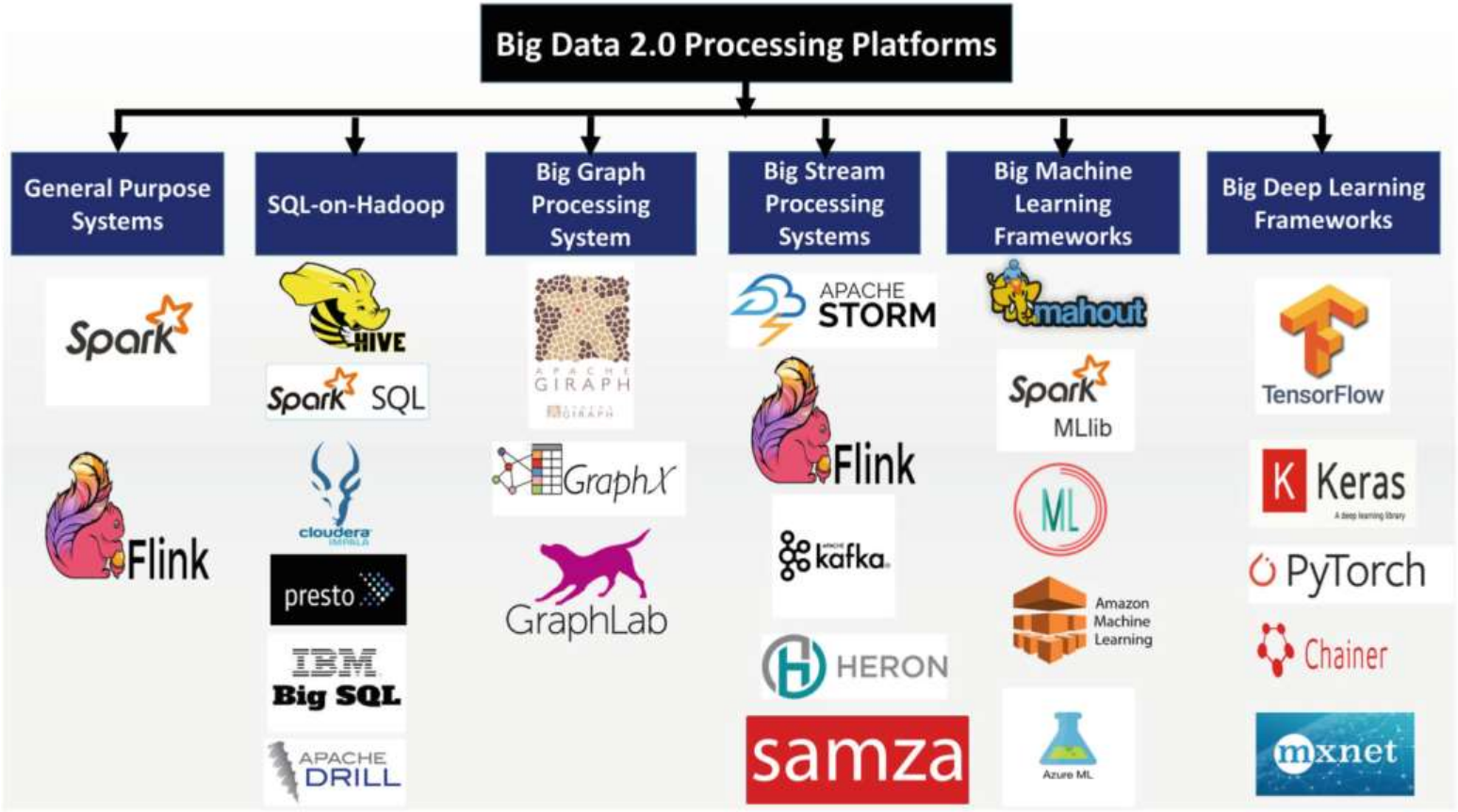
The Process steps of Lambda Architecture



An Example of Implementation of Lambda Architecture



Big Data 2.0 processing systems



Source: Sherif Sakr, *Big Data 2.0 Processing Systems: A Survey, 2nd Edition*, Springer, 2020.

Summary

BDA = ML + CC

- Big Data Analytics: the execution of machine learning tasks on large-datasets in cloud computing environments