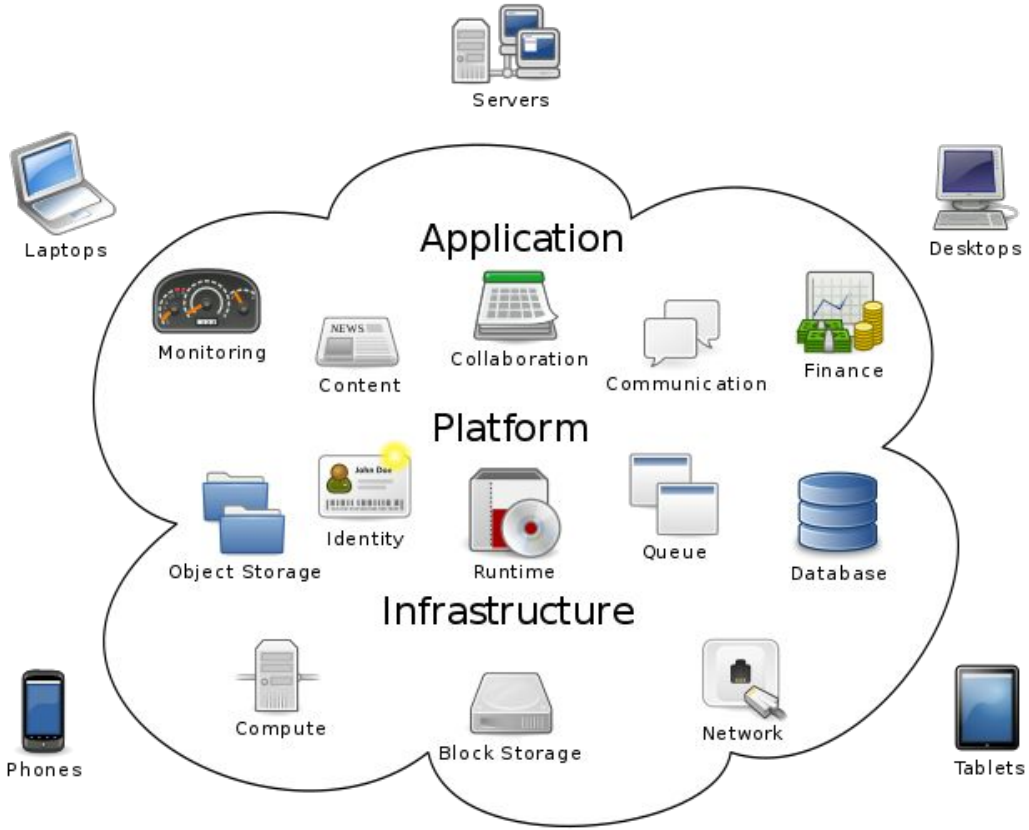
Cloud Computing, Big Data, and the Semantic Web

Dr. V. Raghava Mutharaju
IIIT-Delhi



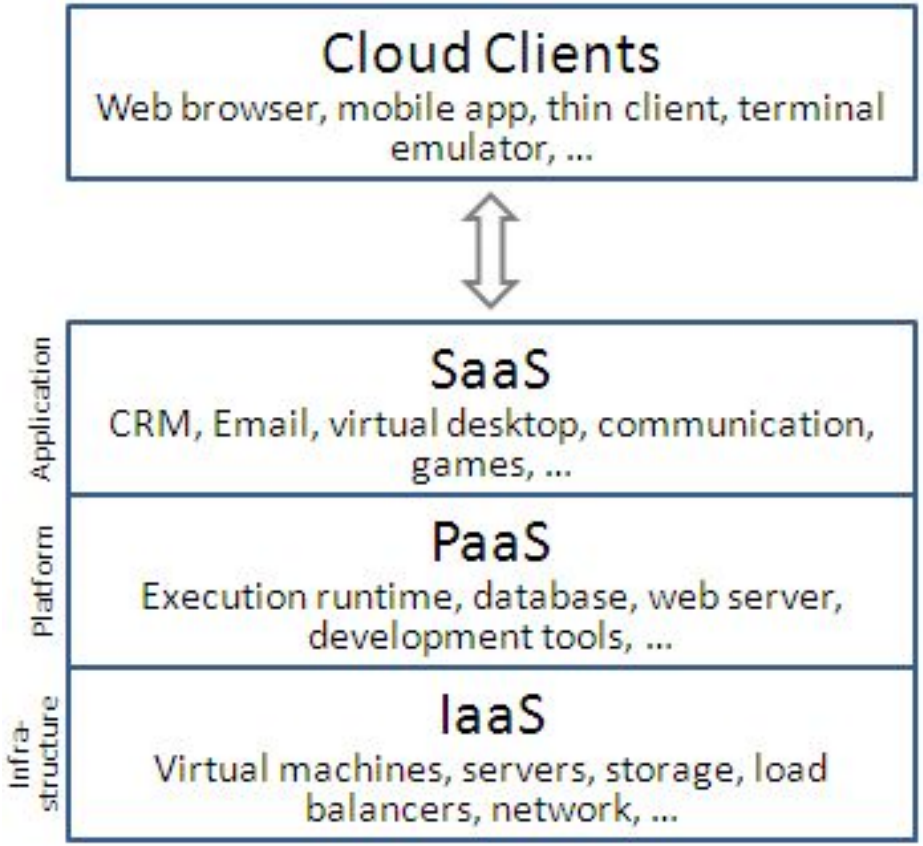
Cloud computing

Cloud Computing

- Laptops → Servers → Cloud
- Maintenance of machines/software
- Availability of machines
- Need the machines/software for only a limited amount of time
- Don’t have the expertise to setup the software on networked machines

- Pay-as-you-go model

Cloud Computing



IAAS

- Infrastructure As A Service (IAAS)
  - Physical computing resources are available as a service in the form of VMs
  - Number of cores
  - RAM size
  - Disk

PAAS

- Platform As A Service (PAAS)
  - Platform/Application environment is provided as a service
  - Operating System
  - Database
  - Java runtime
  - .NET runtime

SAAS

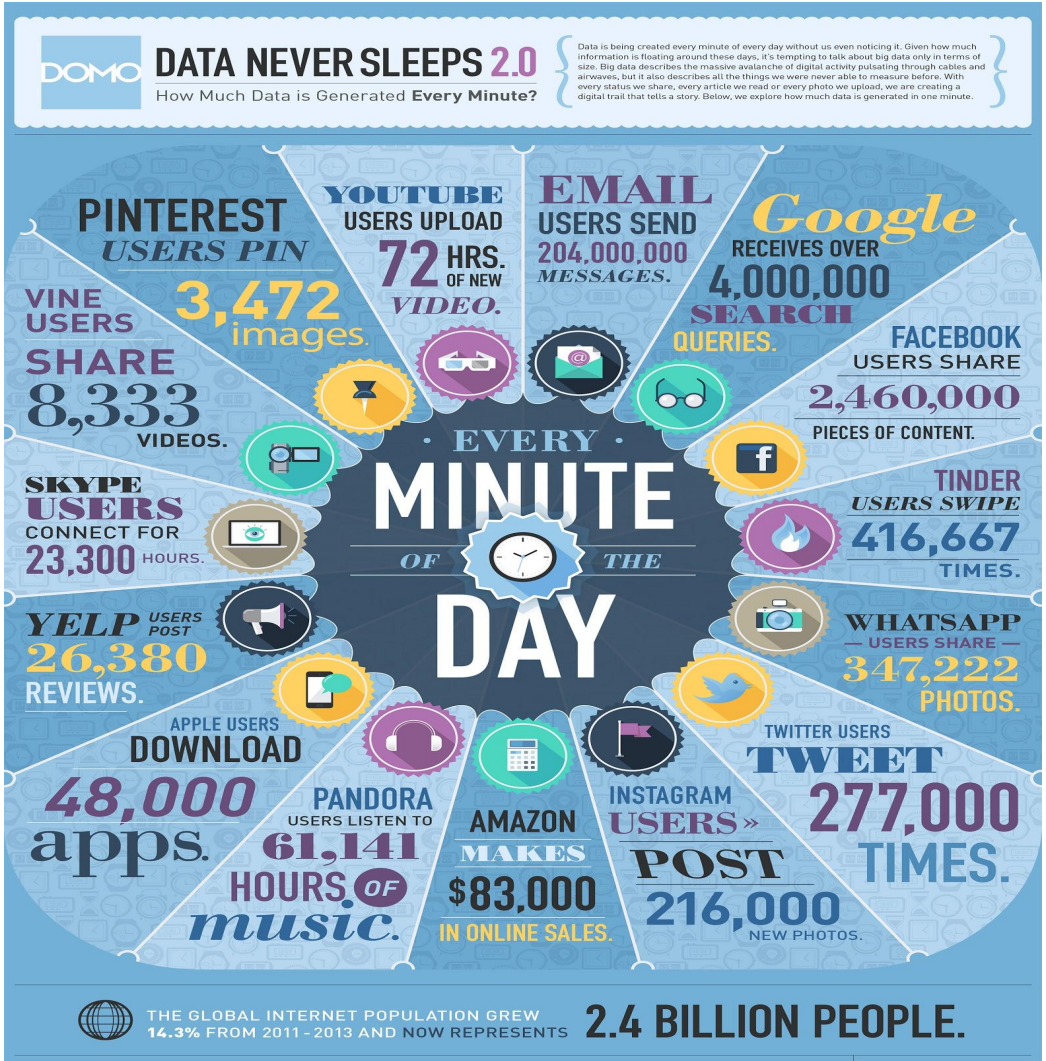
- Software As A Service (SAAS)
  - Software is provided as a service
  - Office software
  - Messaging software
  - CAD software
  - Payroll, Accounting software
  - Content management software

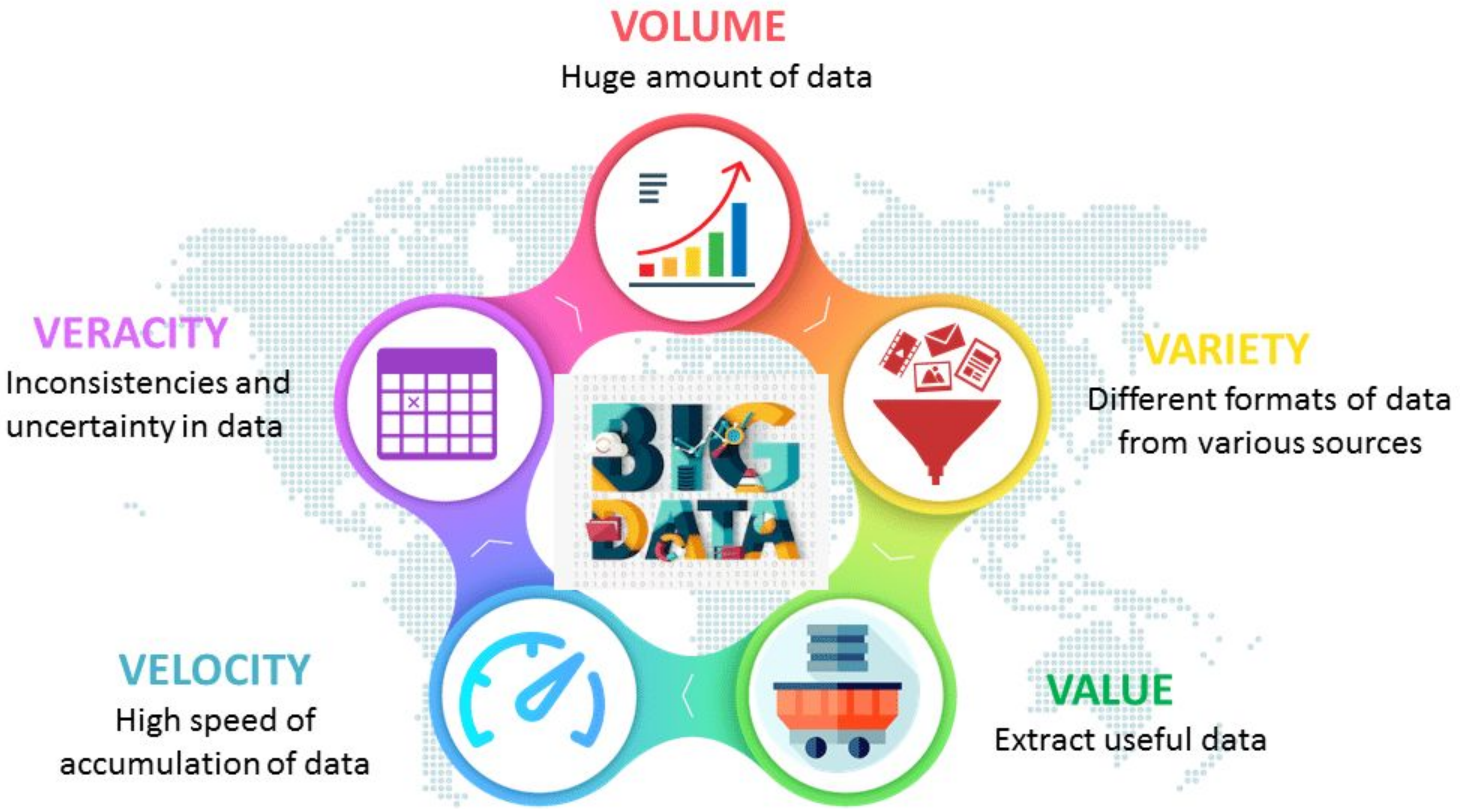
IBM Cloud

- Combines PAAS with IAAS

- Demo

Big Data


DOMO
DATA NEVER SLEEPS 2.0
How Much Data is Generated Every Minute?
Data is being created every minute of every day without us even noticing it. Given how much information is floating around these days, it's tempting to talk about big data only in terms of size. Big data describes the massive avalanche of digital activity pulsating through cables and airwaves, but it also describes all the things we were never able to measure before. With every status we share, every article we read or every photo we upload, we are creating a digital trail that tells a story. Below, we explore how much data is generated in one minute.
PINTEREST USERS PIN 3,472 images.
YOUTUBE USERS UPLOAD 72 HRS. OF NEW VIDEO.
EMAIL USERS SEND 204,000,000 MESSAGES.
Google RECEIVES OVER 4,000,000 SEARCH QUERIES.
FACEBOOK USERS SHARE 2,460,000 PIECES OF CONTENT.
VINE USERS SHARE 8,333 VIDEOS.
SKYPE USERS CONNECT FOR 23,300 HOURS.
EVERY MINUTE OF THE DAY
TINDER USERS SWIPE 416,667 TIMES.
YELP USERS POST 26,380 REVIEWS.
WHATSAPP USERS SHARE 347,222 PHOTOS.
APPLE USERS DOWNLOAD 48,000 apps.
PANDORA USERS LISTEN TO 61,141 HOURS OF music.
AMAZON MAKES $83,000 IN ONLINE SALES.
INSTAGRAM USERS POST 216,000 NEW PHOTOS.
TWITTER USERS TWEET 277,000 TIMES.
THE GLOBAL INTERNET POPULATION GREW 14.3% FROM 2011 - 2013 AND NOW REPRESENTS 2.4 BILLION PEOPLE.


VOLUME
Huge amount of data
VERACITY
Inconsistencies and uncertainty in data
VARIETY
Different formats of data from various sources
BIG DATA
VELOCITY
High speed of accumulation of data
VALUE
Extract useful data

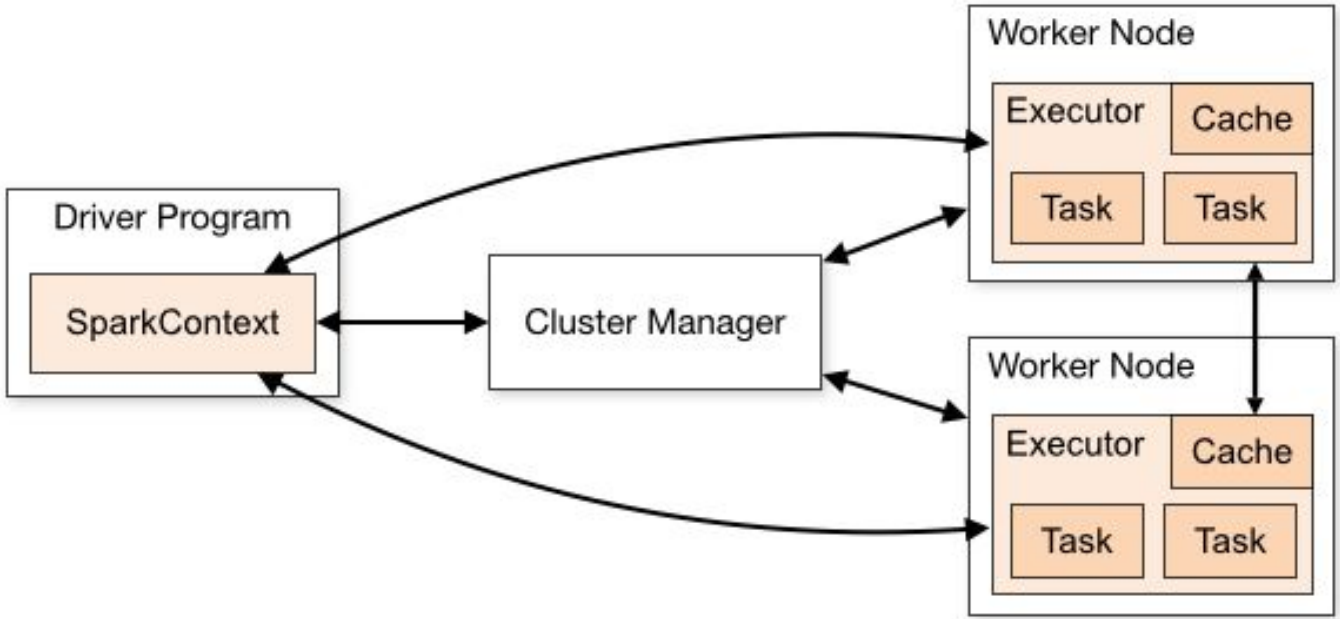
Apache Spark

- Handles the volume aspect very well
- Has also support for velocity
- Has support for Graph and Machine Learning algorithms that can be used to generate value from the data

Apache Spark

- Distribute the data and the computations across a cluster (group of machines)
- This increases the parallelism and reduces the load on each machine
  - Need more machines? Pay-as-you-go (add more machines depending on the load)
- Spark provides mechanism to
  - Capture the data in a form that is easily distributable
  - Perform computations that can be parallelized

Spark Cluster



Spark RDD

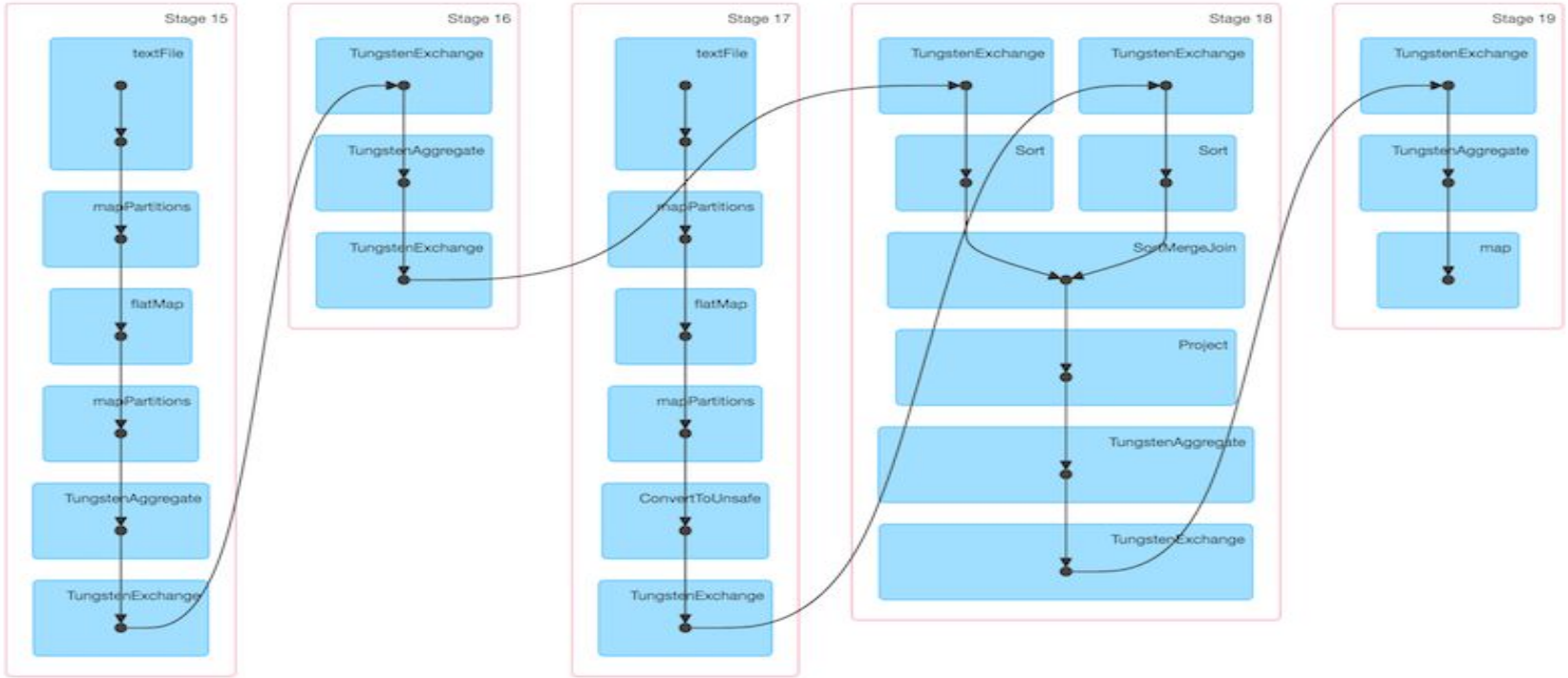
- Resilient Distributed Dataset (RDD)
  - It is the fundamental datastructure of Apache Spark
  - It is a immutable, fault tolerant collection of elements that can be operated in parallel
  - An RDD can be divided into logical partitions that can be distributed across the cluster
  - Elements of RDD can be text, numbers, or a combination of them

Transformations and Actions

Transformations				
map	join	union	distinct	repartition
mapPartitions	flatMap	intersection	pipe	coalesce
cartesian	cogroup	filter	sample	
sortByKey	groupByKey	reduceByKey	aggregateByKey	
mapPartitionswithIndex		repartitionAndSortWithinPartitions		

Actions				
reduce	take	collect	takeSample	count
takeOrdered	countByKey	first	foreach	saveAsTextFile
saveAsSequenceFile		saveAsObjectFile		

Directed Acyclic Graphs



Big Data Problems

- Parallelizing very large datasets (Protein sequencing)
- Estimate the runtime of Apache Spark jobs
- Automate the generation of code based on text descriptions and heuristics