Hadoop and Spark for Data Scientists

Lecture 1 : Distributed Systems, Big Data and Hadoop Ecosystem

(Data Storage and Processing at Scale)

Hassan Ahmadkhani

dayche.com . گروه دایکه

DAYCHE دای‌کِه

رئوس مطالب

- مفاهیم کلان داده، ذخیره سازی توزیع شده
- معرفی و کار با Hadoop
- پیاده سازی انبار داده در هادوپ به کمک Apache Hive
- پایگاه داده های NoSQL و کار با MongoDB و Elastic
- معرفی اسپارک و توسعه App به کمک Spark Core
- عملیات داده ای به کمک Spark SQL

تولید محتوا: حسن احمدخانی
daychegroup
daychegroup
dayche.com | گروه دایکه

انواع داده ها

- Relational Data (Tables/Transaction)
- Text Data (Web)
- Image Data
- Video Data
- Audio Data
- Semi-structured Data (XML, JSON, …)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once

❑ What is Big Data?

No single definition; here is from Wikipedia and IBM:

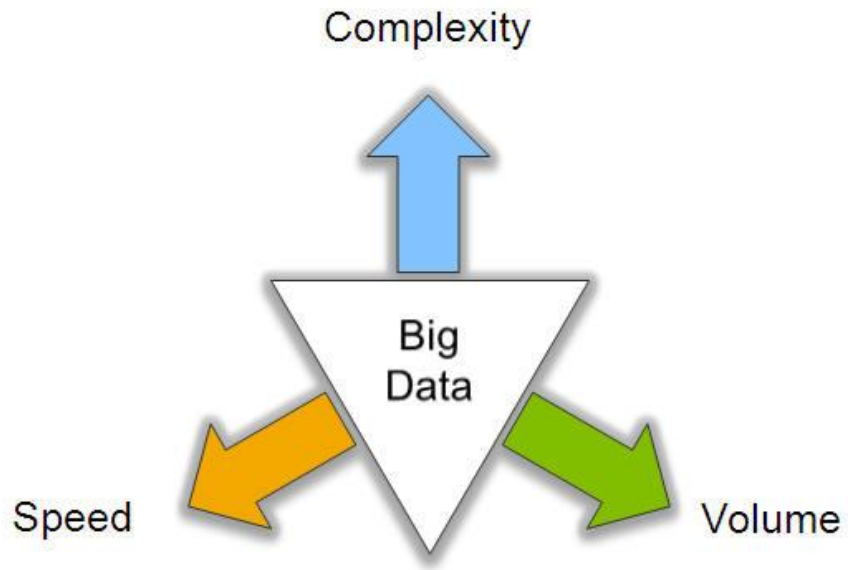
Big data is the term for a collection of data sets so large(size) and complex(type) that it becomes difficult(beyond the ability) to use:

- relational databases to capture, manage, and process the data
- or traditional data processing applications.

with low-latency

The challenges include:

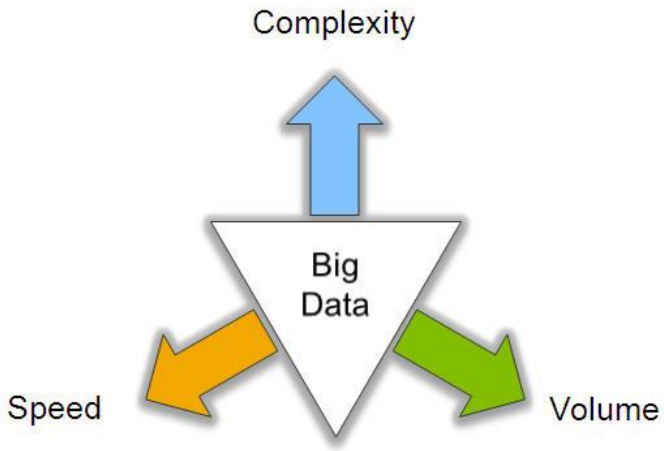
- Capture
- Storage
- Search
- Sharing
- Transfer
- Analysis
- and Visualization



❑ What is Big Data?

Numbers: sample for relational data (data without comple

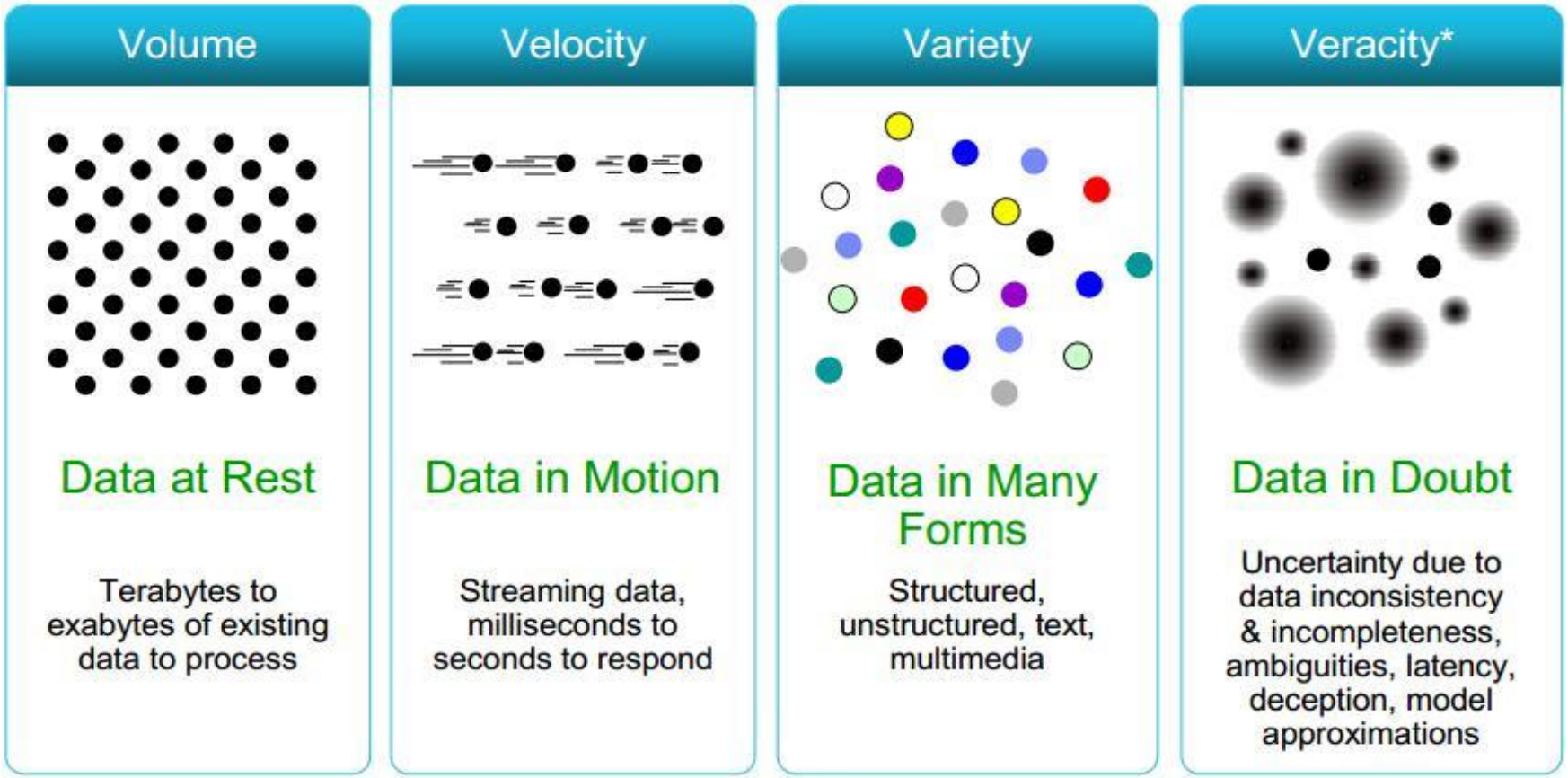
- 100M Record …
- 500M Record …
- 60B Record …
- 50 TB Table Size
- 100+TB Table Size (daily 1B Record …)
- …



Big Data for Data Scientists

- ML Workloads on Large Datasets

❑ What is Big Data? Some Make it 4V's



Distributed Processing and Storage

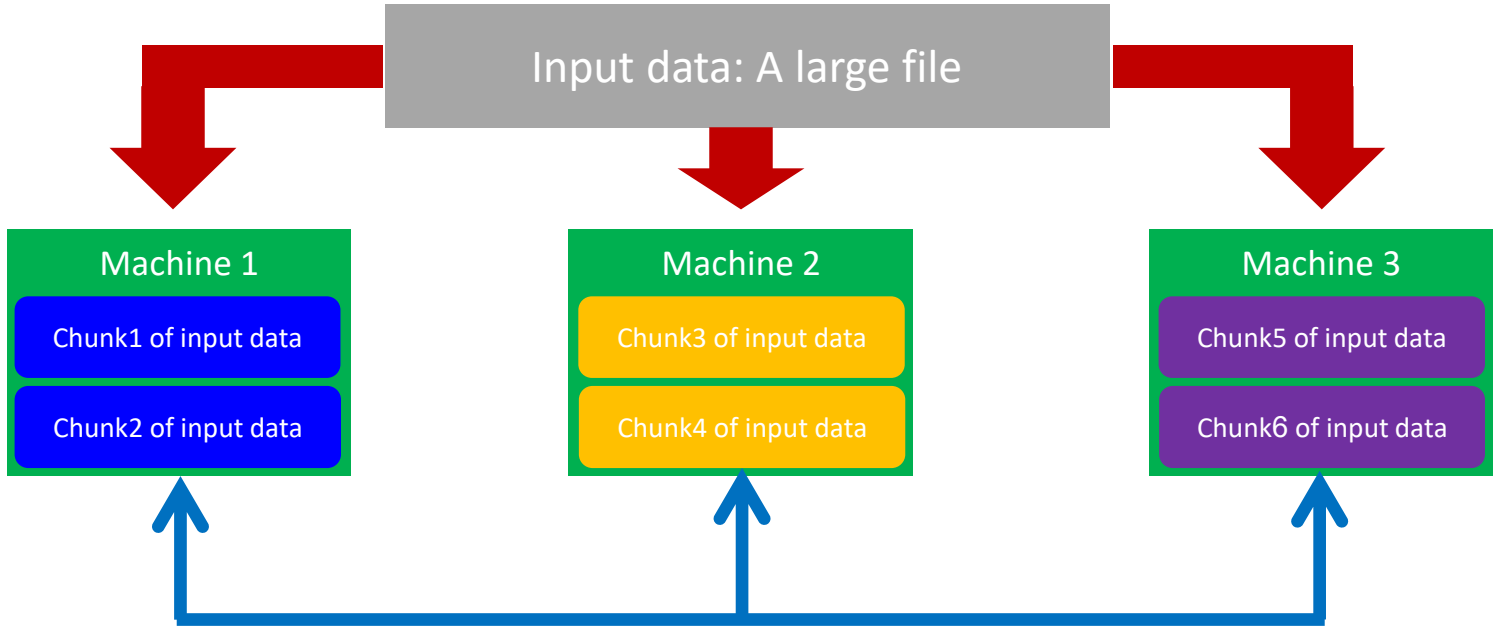
Manage multiple machines as if they were a single computer

- پردازش موازی
- پردازش توزیع شده
- ذخیره سازی توزیع شده

تولید محتوا: حسن احمدخانی
daychegroup
daychegroup
dayche.com | گروه دایکه

Why Sharding or Data Distribution ?

- Data is typically sharded (or striped) to allow for concurrent accesses



Scaling Databases

❑ **RDBMSs can be either scaled:**

- **Vertically (or Up)**
  - Can be achieved by hardware upgrades (e.g., faster CPU, more memory, or larger disk)
  - Limited by the amount of CPU, RAM and disk that can be configured on a single machine

- **Horizontally (or Out)**
  - Can be achieved by adding more machines
  - Requires database *sharding* and probably *replication*
  - Limited by the Read-to-Write ratio and communication overhead

❑ مروری بر چند مفهوم

- DW
- OLTP
- ACID Properties
- OLAP
- ETL
- ELT
- FT - تحمل پذیری خطا
  - سخت افزاری
  - نرم افزاری
- Scalability – مقیاس پذیری
- Data Scan – Table Scan
- Random Access

❑ Data Lake

❑ Data Lakehouse

تولید محتوا: حسن احمدخانی
daychegroup
daychegroup
dayche.com | گروه دایکه

Hadoop - definition from cloudera:

- **Apache Hadoop** is a software platform for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware.

- **Hadoop services provide for:**
  - data storage,
  - data processing,
  - data access,
  - data governance,
  - security,
  - and operations.

The Big Data Marketplace

- **Hadoop distributions** – Cloudera and Hortonworks, MapR - HP, Open Source Edition

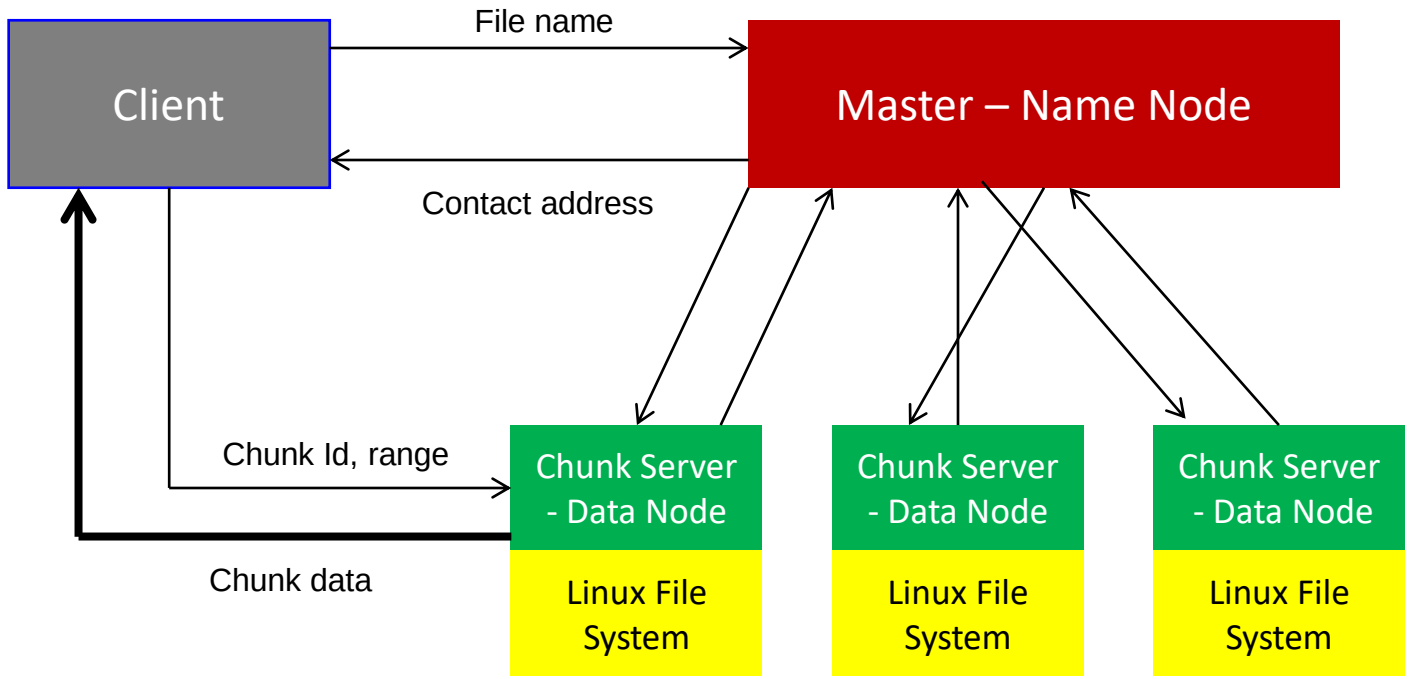
- **Big Data Appliances and Cloud Solutions :**
  - GCP , AWS , …
  - Oracle Big Data Appliance,
  - IBM BigInsights,
  - Microsoft PDW and HD Insight,
  - EMC GreenPlum DCA
  - PivotalHD

Hadoop Components

- Hadoop Distributed File System (HDFS)
  - Single namespace for entire cluster
  - Replicates data for fault-tolerance
- Distributed Resource Management (YARN)
- Distributed Processing Framework (MapReduce)

HDFS Architecture

- HDFS adopts a master-slave architecture
- HDFS Main Components

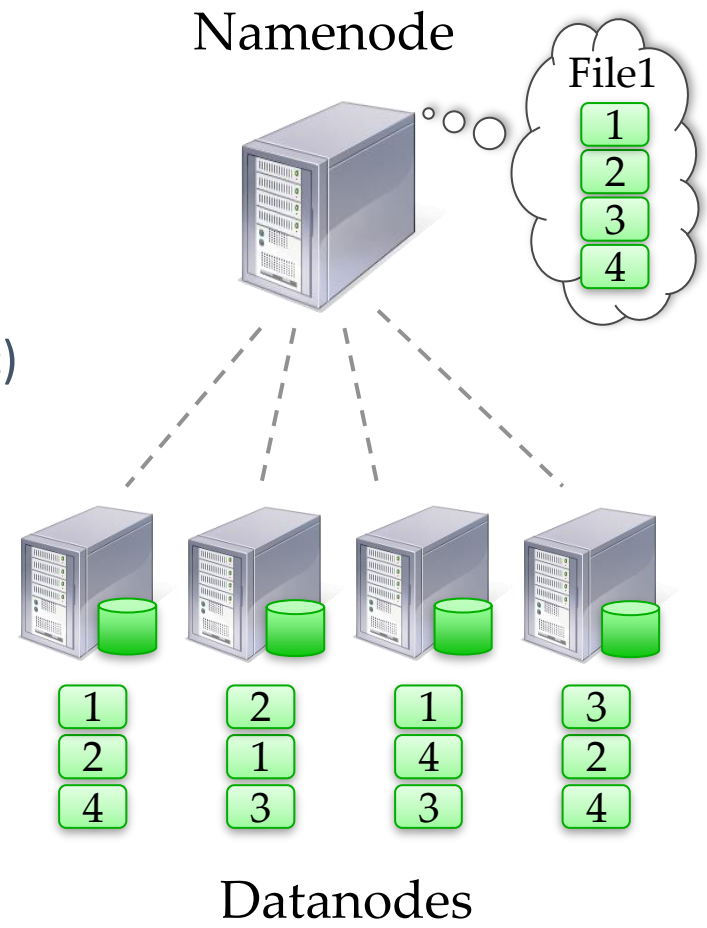


HDFS Data Distribution Policy

- The Hadoop Distributed File System (HDFS) is a scalable DFS for data-intensive applications
- HDFS divides large files into multiple pieces called chunks or blocks (by default 128MB) and stores them on different data servers
  - This design is referred to as block-based design
- Each HDFS chunk has a unique 64-bit identifier and is stored as a file in the lower-layer local file system on the data server
- HDFS distributes chunks across cluster data servers using a random distribution policy

HDFS Random Distribution Policy

- Files split into 128MB blocks
- Blocks replicated across several datanodes (often 3)
- Namenode stores metadata (file names, locations, etc)
- Optimized for large files, sequential reads
- Files are append-only



Hadoop - YARN

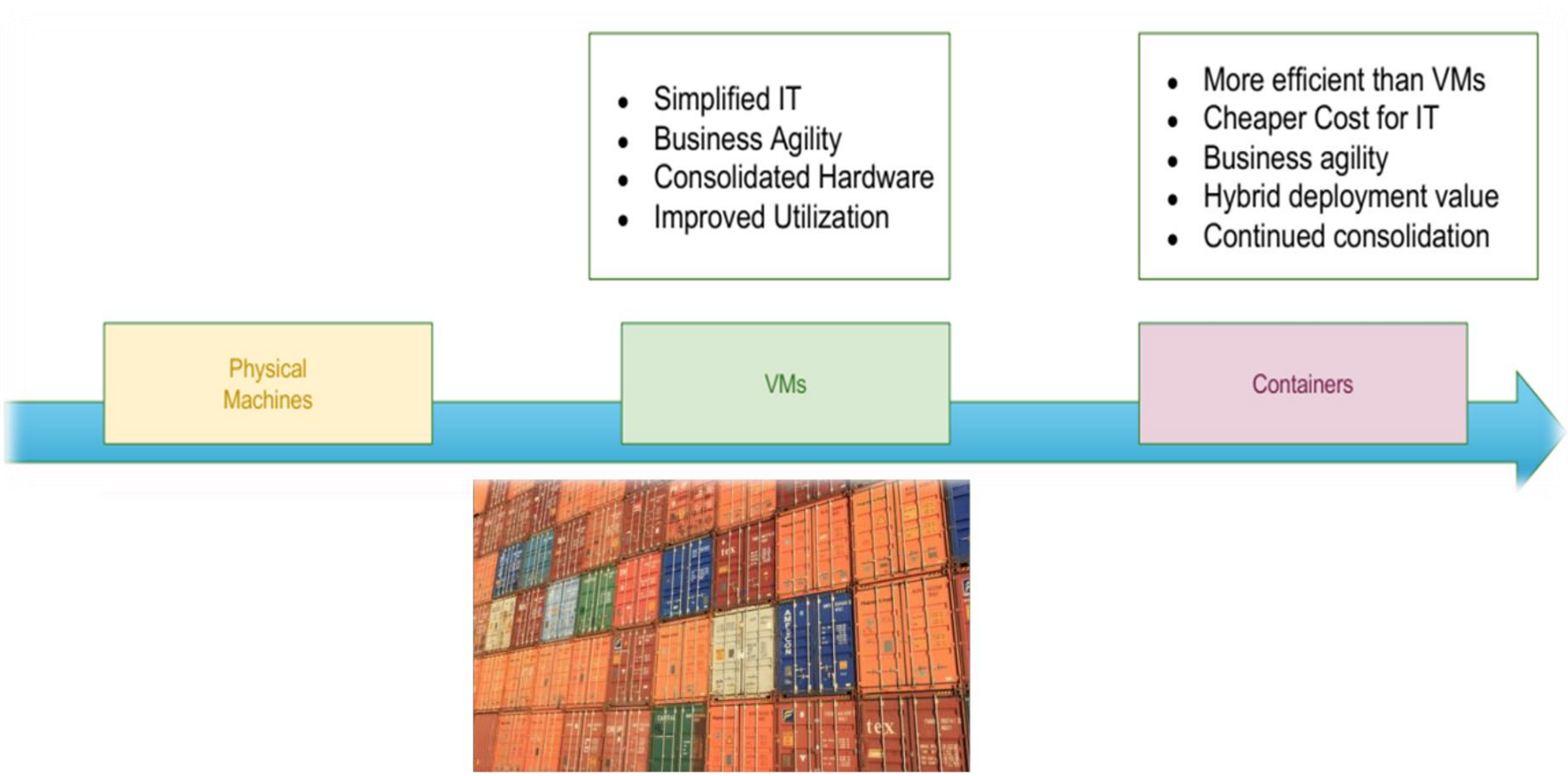
- Apache Hadoop is software platform for distributed storage and distributed processing (resource and job management) of very large data sets on computer clusters built from commodity hardware.
- Hadoop services provide for
  - data storage,
  - data processing,
  - data access,
  - data governance,
  - security,
  - and operations.
- Hadoop Distributed Resource and Job Manager
  - Yet Another Resource Negotiator (YARN)

YARN Architecture

- Fundamental idea of YARN is to split up the functionalities
  - Resource management
  - Job scheduling/monitoring
  - into separate daemons.

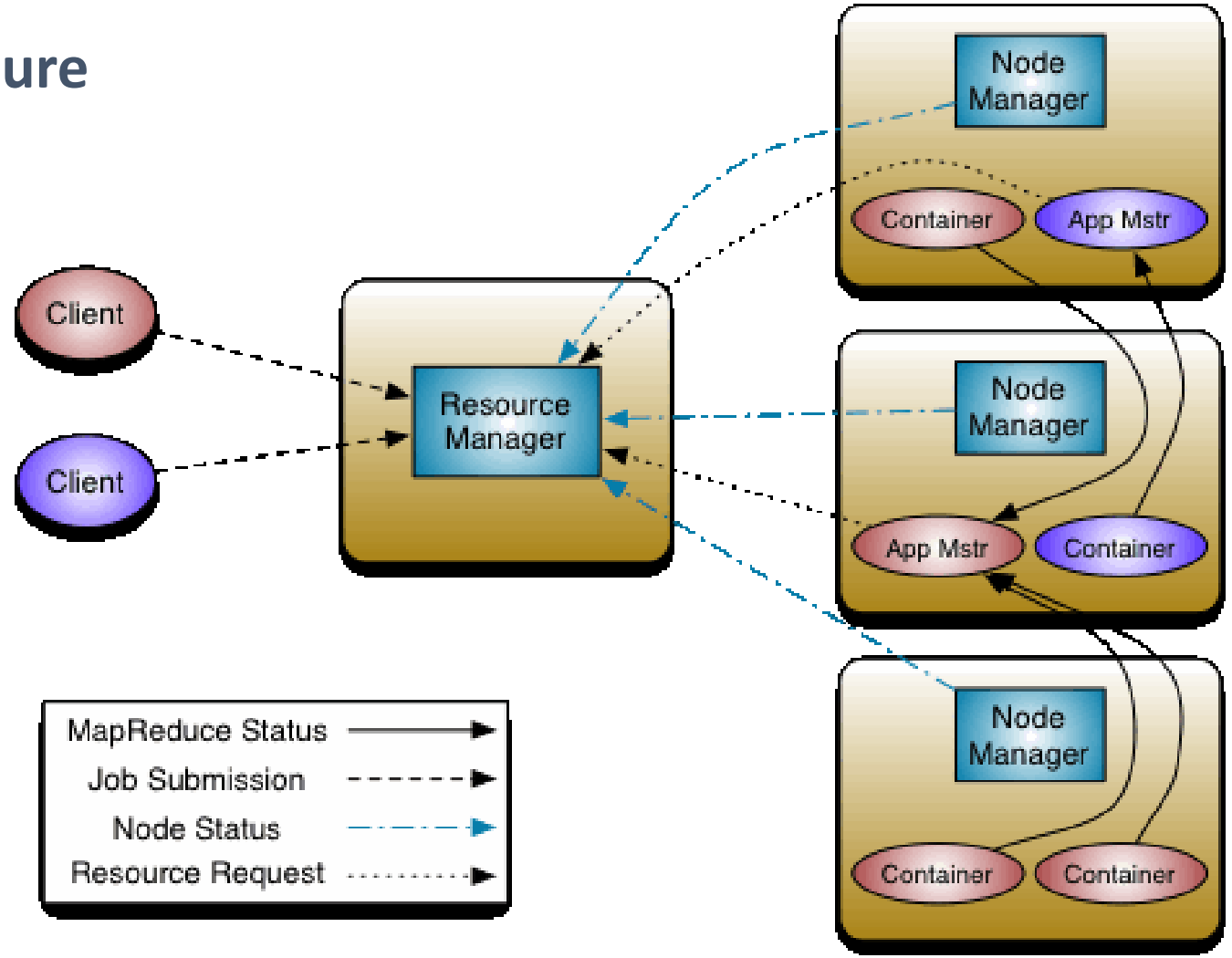
- The idea is to have a global Resource Manager (RM) and per-application Application Master (AM)
- An application is either a single job or a DAG of jobs
- **YARN main components:**
  - ResourceManager : is the ultimate authority that arbitrates resources among all the applications in the system
  - NodeManager : per-machine agent who is responsible for containers, monitoring their resource usage (cpu, memory, disk, network) and reporting the same to the ResourceManager/Scheduler

YARN Architecture : Containers



تولید محتوا: حسن احمدخانی
daychegroup
daychegroup
dayche.com | گروه دایکه

YARN Architecture

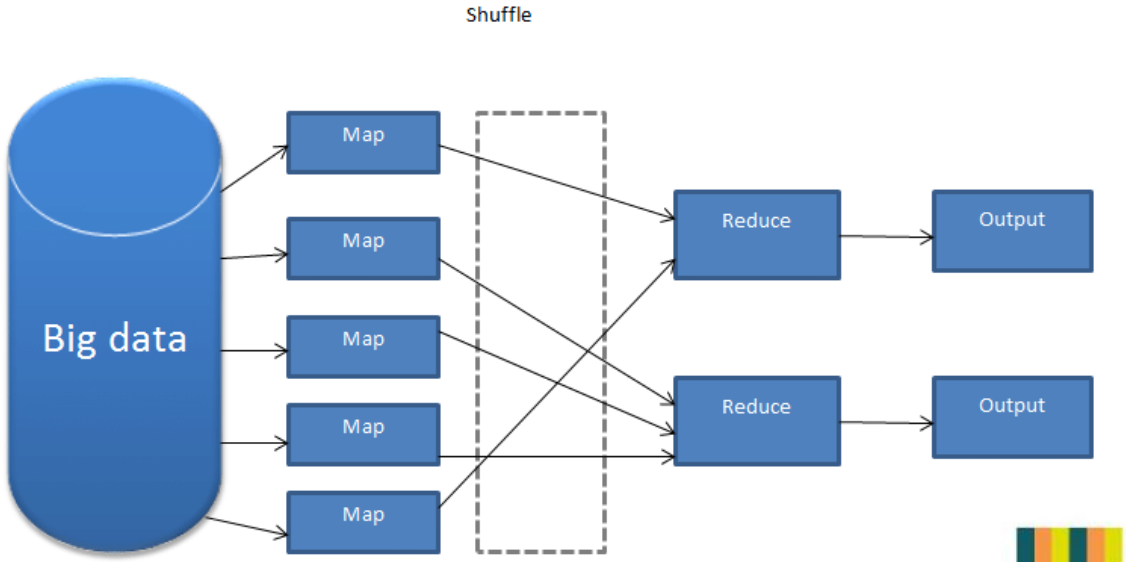


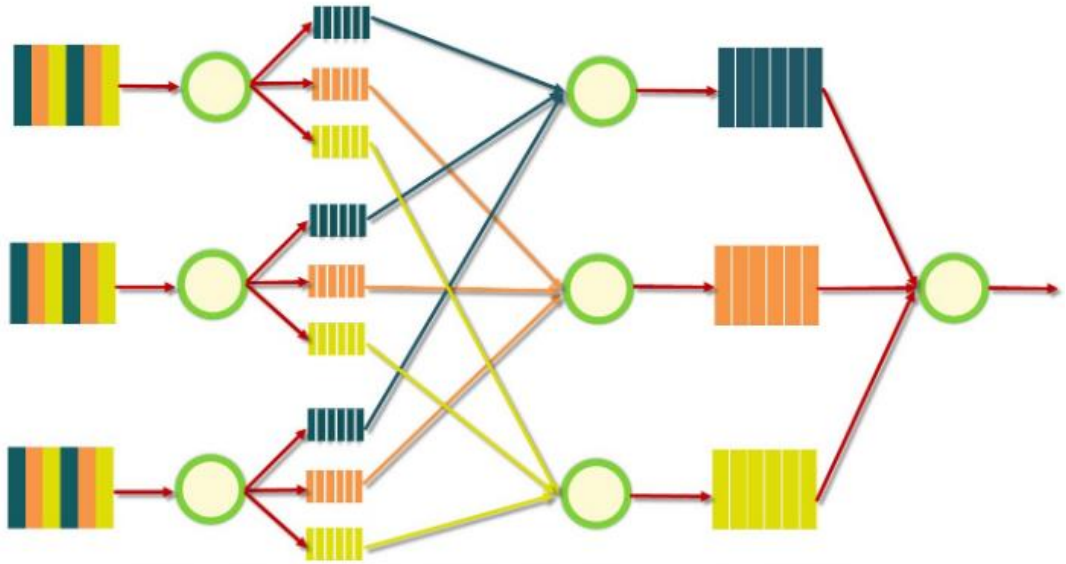
تولید محتوا: حسن احمدخانی
daychegroup
daychegroup
dayche.com | گروه دایکه

Hadoop

- Apache Hadoop is software platform for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware.
- Hadoop services provide for
  - data storage,
  - data processing,
  - data access,
  - data governance,
  - security,
  - and operations.
- Hadoop Distributed Processing Engine
  - Map Reduce (MR)

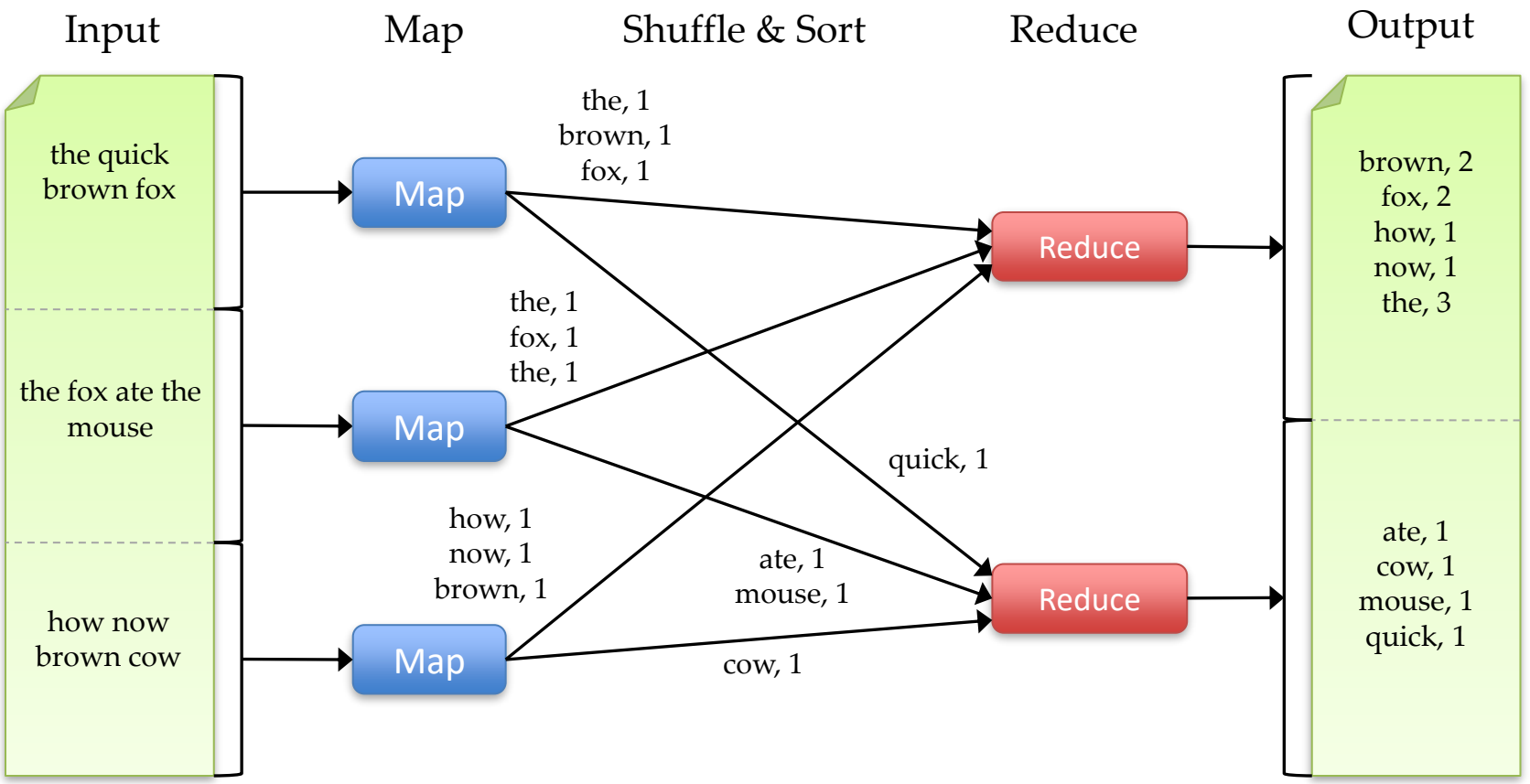
نمایش شماتیک مدل پردازشی Map Reduce



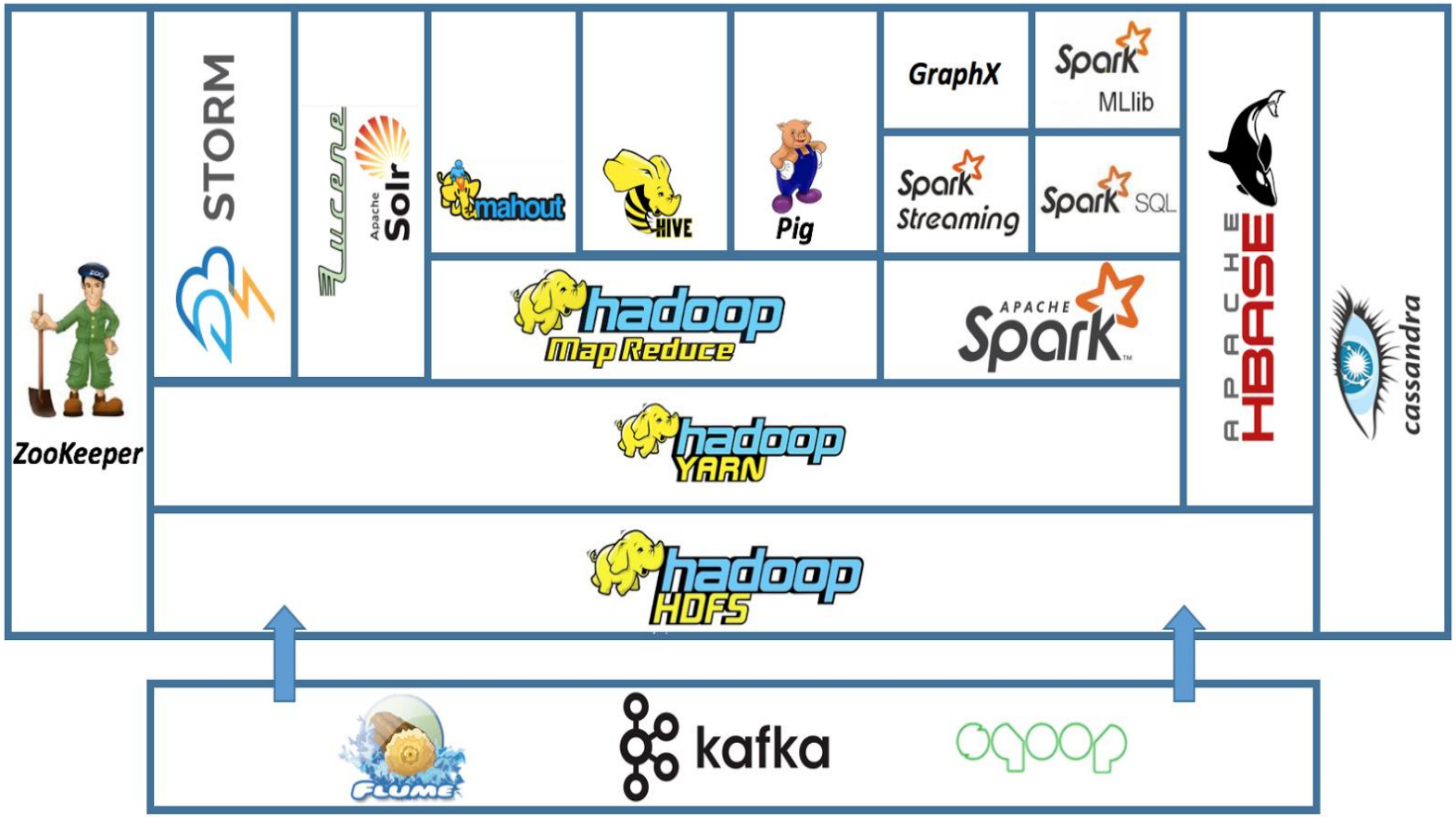


تولید محتوا: حسن احمدخانی
daychegroup
daychegroup
dayche.com | گروه دایکه

معرفی مدل پردازشی نگاشت کاهش / Map Reduce



Hadoop Ecosystem



تولید محتوا: حسن احمدخانی
daychegroup
daychegroup
dayche.com | گروه دایکه