Dr Amin Karami (PhD, FHEA)

a.karami@uel.ac.uk

amin.karami@ymail.com

www.aminkarami.com

What is Big Data?

What is Big Data?

- Big data is an industry term, when they faced with new data problems. Organizations spend millions of dollars on data storages to collect massive data, **and the problem is failing to do anything with them.**

- Organizations need an effective data analytics solutions for [1] **data storage** and [2] the ability to analyse them in **near real time,** with **low latency**.

How to provide an effective solutions?

- Data is generated in many ways. The big question is how/where to put it all together to create value? This challenge can be summarized into four key components/groups (**4V's**): `volume, velocity, variety,` and `veracity`.

- Another big question is how to analyse these different format/group of data? We may require different services/tools/approaches.

Volume

Symbol	Abbr.	Value (byte)	Example
Bit	b	10^0	Single binary digit (0 or 1)
Byte	B	10^1	8 bits = 2^3
Kilobyte	KB	10^3	1,024 B = 2^{10}
Megabyte	MB	10^6 (Million)	1,024 KB = 2^{20}
Gigabyte	GB	10^9 (Billion)	1,024 MB = 2^{30}
Terabyte	TB	10^{12} (Trillion)	1,024 GB = 2^{40}
Petabyte	PB	10^{15} (Quadrillion)	1,024 TB = 2^{50}
Exabyte	EB	10^{18} (Quintillion)	1,024 PB = 2^{60}
Zettabyte	ZB	10^{21} (Sextillion)	1,024 EB = 2^{70}
Yottabyte	YB	10^{24} (Septillion)	1,024 ZB = 2^{80}

Velocity

- When businesses need rapid insights/analytics from the data they are collecting, the systems in place cannot meet the need, there is a velocity problem. This is because of the limited RAM and CPU.
- The data is being generated is accelerating in some applications, such as emails, photos, Twitters, Facebook posts, log files, IoT devices that are being generated rapidly and must be collected, processed, analysed and stored at high speeds. How?

Velocity: Collecting Data

- **Batch and Periodic:** This is an amount of data that can be transferred/collected at scheduled intervals. We have well enough time to collect data and plan for the appropriate resources.

- **Near real-time:** velocity is a huge concern with near real-time processing. These systems require data to be collected and processed within minutes. We have no time to collect large size of data, we can process small size of data and plan accordingly in terms of the availability of CPU and RAM.

- **Real-time:** velocity is the paramount concern. Data canNOT take minutes to process. The processes should be done in seconds and maintain the useful information/data.

Velocity: Batch vs Stream processing

	Batch data processing	Streaming data processing
Data Scope	Processing over all or most of the data	Processing over data within a rolling time window, or the most recent data
Data Size	Large batches of data	Individual records or micro batches consisting of a few records
Latency	Minutes to hours	Seconds to milliseconds
Analysis	Complex Analysis	Simple response functions, aggregates, and rolling metrics

Variety

- Data comes from different sources and this wide variety becomes a challenge for businesses facing with diversity in the analytics.

- **Structured Data:** it is stored in a tabular format: RDBMS. Such as, Microsoft SQL, MySQL, PostgreSQL, Amazon RDS, Oracle, etc.

- **Unstructured Data:** it is stored in the form of files. It can be texts, photos, audio records, videos, clickstream data, Amazon S3, Amazon Redshift Spectrum and many more. It consists of irrelevant information, not like relational databases.

Variety

- **Semi-structured Data:** it is stored in the form of elements within a file. The data is organized based on the elements and the attributes and they have a self-describing structure. The attributes within an element define the characteristics of the file. this makes semi-structured data flexible and able to scale to meet the changing demands of a business much more quickly that structured data. However, it can be more difficult to analyse them when analysts cannot predict which attributes will be present in ant given data set. They include CSV, XML, JSON, Amazon DynamoDB etc.

Veracity

- It refers to the quality of the data that is being analysed. With so much data, ensuring it is relevant, accuracy, not-duplicated, noise-free, abnormality-free, completeness is the big problem for consideration.

Knowledge Check

Scenario 1

My business has a set of 15 JSON data files that are each about 2.5 GB in size. They are placed on a file server once an hour. They must be ingested as soon as they arrive in this location. This data must be combined with all transactions from the financial dashboard for this same period, then compared to the recommendations from the marketing engine. All data is fully cleansed. The results from this time period must be made available to decision makers by 10 minutes after the hour in the form of financial dashboards.

Source: Amazon AWS

Based on the scenario, which of the following Vs pose a challenge for this business?

- ❑ Volume
- ❑ Velocity
- ❑ Variety
- ❑ Veracity

Knowledge Check

Scenario 2

My business compiles data generated by hundreds of corporations. This data is delivered to us in very large files, transactional updates, and even data streams. The data must be cleansed and prepared to ensure that rogue inputs do not skew the results. Knowing the data source for each record is vital to the work we do. A large portion of the data gathered is irrelevant to our analysis, so this data must be eliminated. The final requirement is that all data must be combined and loaded into our data warehouse, where it will be analysed.

Source: Amazon AWS

Based on the scenario, which of the following Vs pose a challenge for this business?

- ❑ Volume
- ❑ Velocity
- ❑ Variety
- ❑ Veracity

Why do we need a new tech?

Data: 5 TB
Speed of machine: 200 MB/sec

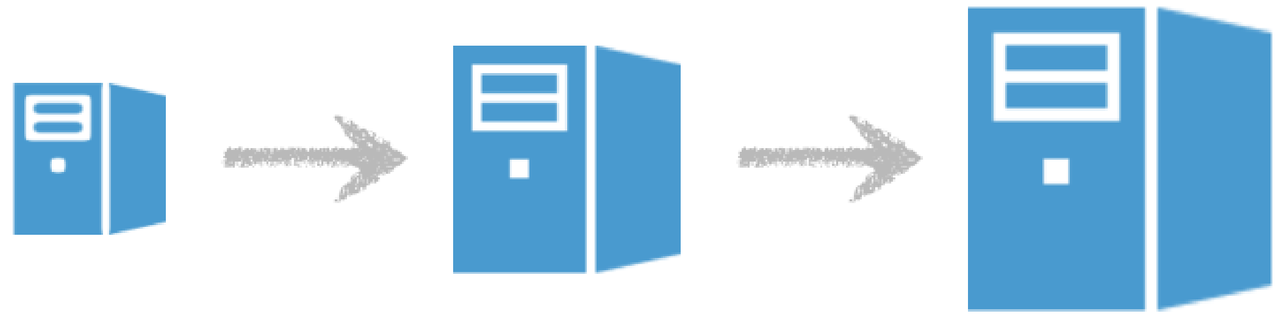
How much time will take to process 5 TB data?

$$Processing = \frac{5 * 1024 * 1024}{200 * 60} = approx.\ 437\ Mins$$

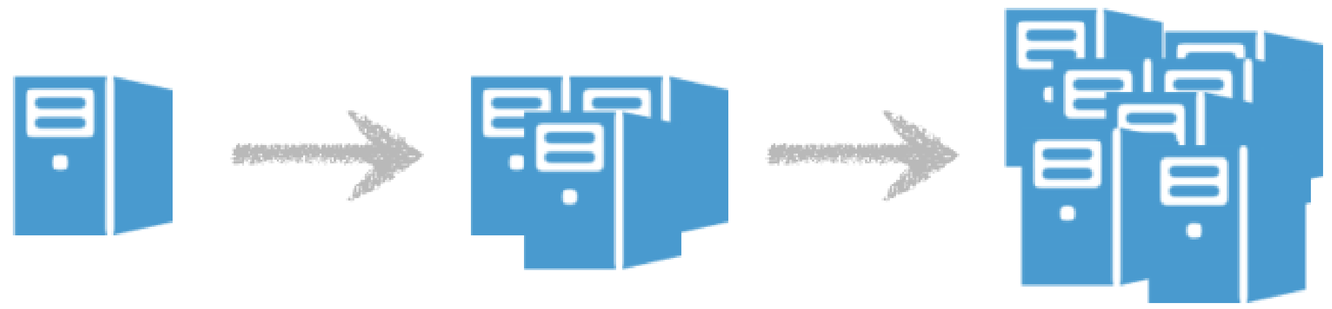
Do you think does it reasonable? Could we process our data?
Certainly, NO

How to process Big Data?

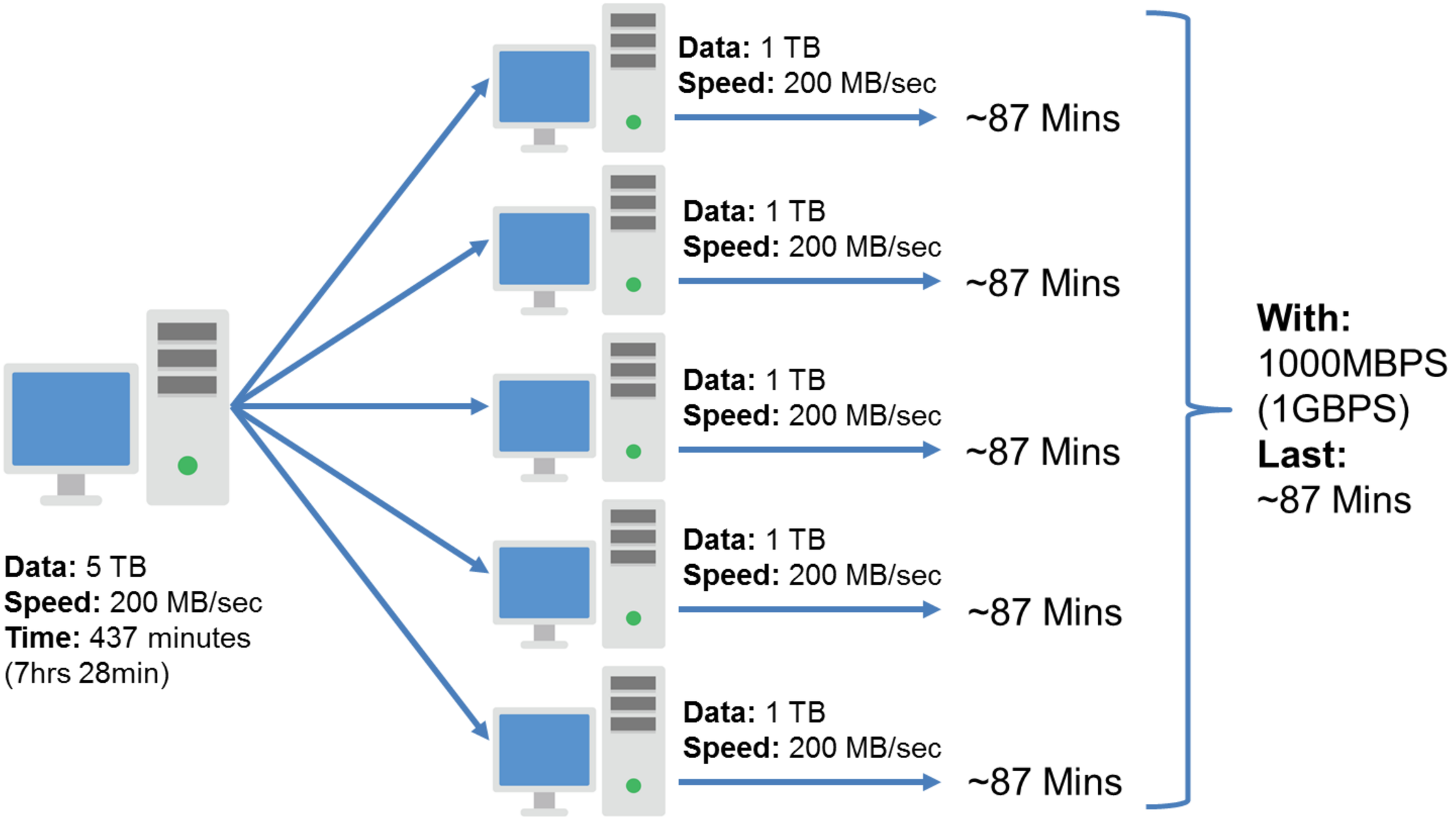
- **Scale Up** (scale vertically): adding resources to a single node in a system. [expensive, not possible for massive data]



- **Scale Out** (scale horizontally): adding more nodes to a system. [expensive, fault tolerance, development problems]



Reading big data: parallel processing



Parallel Processing Challenges

1. **Dividing and Distributing** Divide 5TB data into 1TB and send them into workstations

2. **Parallel Processing** Run all the workstations in parallel without delay and fault

3. **Combining Results** We need combine all the results from each workstation

4. **Costly Servers** more workstations take more cost

How to overcome these challenges?

Apache Hadoop

Source: IBM hadoop

- **Hadoop** was created by computer scientists Doug Cutting and Mike Cafarella in 2006 to support distribution for the **Nutch search engine**. It was inspired by **Google File System** (Oct. 2003) **Google MapReduce** (Dec. 2004), a software framework in which an application is broken down into numerous small part.

What is Hadoop?

- **Apache Hadoop** is an open-source framework written in Java, that supports the processing and storing of massive amount of data (referred as Big Data) on clusters of commodity hardware.

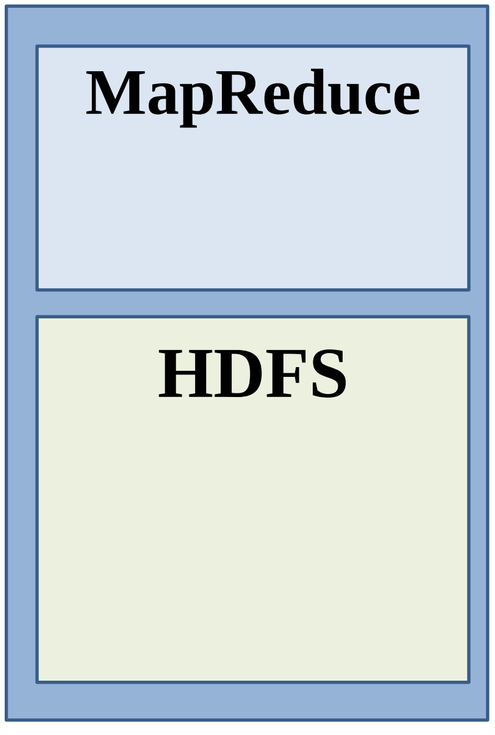
- **Apache Hadoop 1.0** became publically available in **November 2012** as part of the Apache project sponsored by the Apache Software Foundation.

- The latest stable release is available in Apache website: https://hadoop.apache.org/releases.html

Hadoop Pieces

1. **HDFS** (Hadoop Distributed File System) is the data part of Hadoop which is the primary storage system used by Hadoop applications.

2. **MapReduce** is the processing part of Hadoop.



HDFS components:

- **NameNode (Master)** is a centrepiece of HDFS that stores the metadata of HDFS. It keeps a list of **blocks and its location** for files stored in HDFS. It is a single point of failure. It needs more memory (RAM) to be executed.
- **DataNode (Slave)** is responsible for storing the actual data in HDFS. It is usually configured with a lot of hard disk.

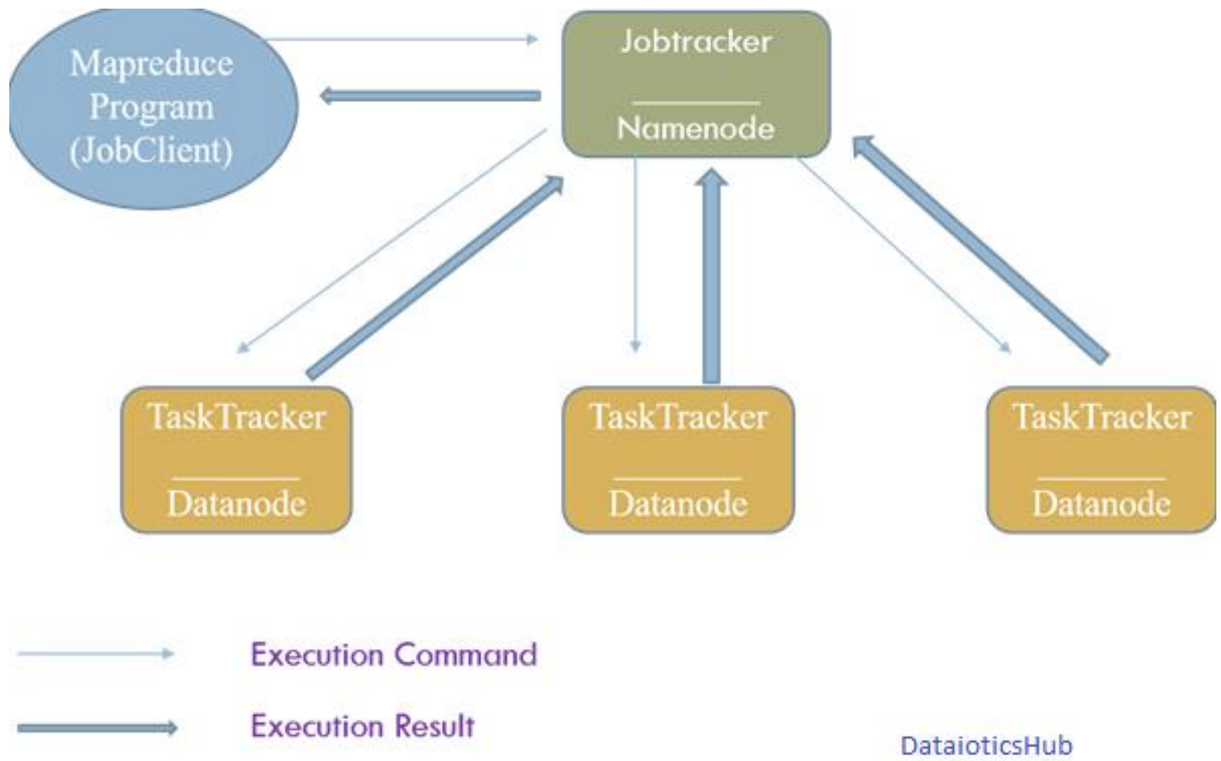
HDFS components

- **NameNode**
  - Keep all namespaces in memory
  - Maintains metadata
  - Monitors data node health
  - Replicates missing blocks
  - Maintains mapping of list of blocks and locations
  - Maintains authorization and authentication data
  - Manages checkpoints
- **DataNode**
  - Serves data blocks directly to clients
  - Handles block storage on multiple volumes, block integrity
  - Periodically sends heartbeats and block reports to NameNode
  - Stores blocks as underlying OS's file

MapReduce components

- **JobTracker**: A coordinator for tasks. It keeps track of jobs being run on clusters. Each cluster has a single job tracker. It settles in the Memory.

- **TaskTracker**: A node that accepts tasks (e.g., **Map, Reduce, Shuffle**). Each cluster can have multiple task trackers.



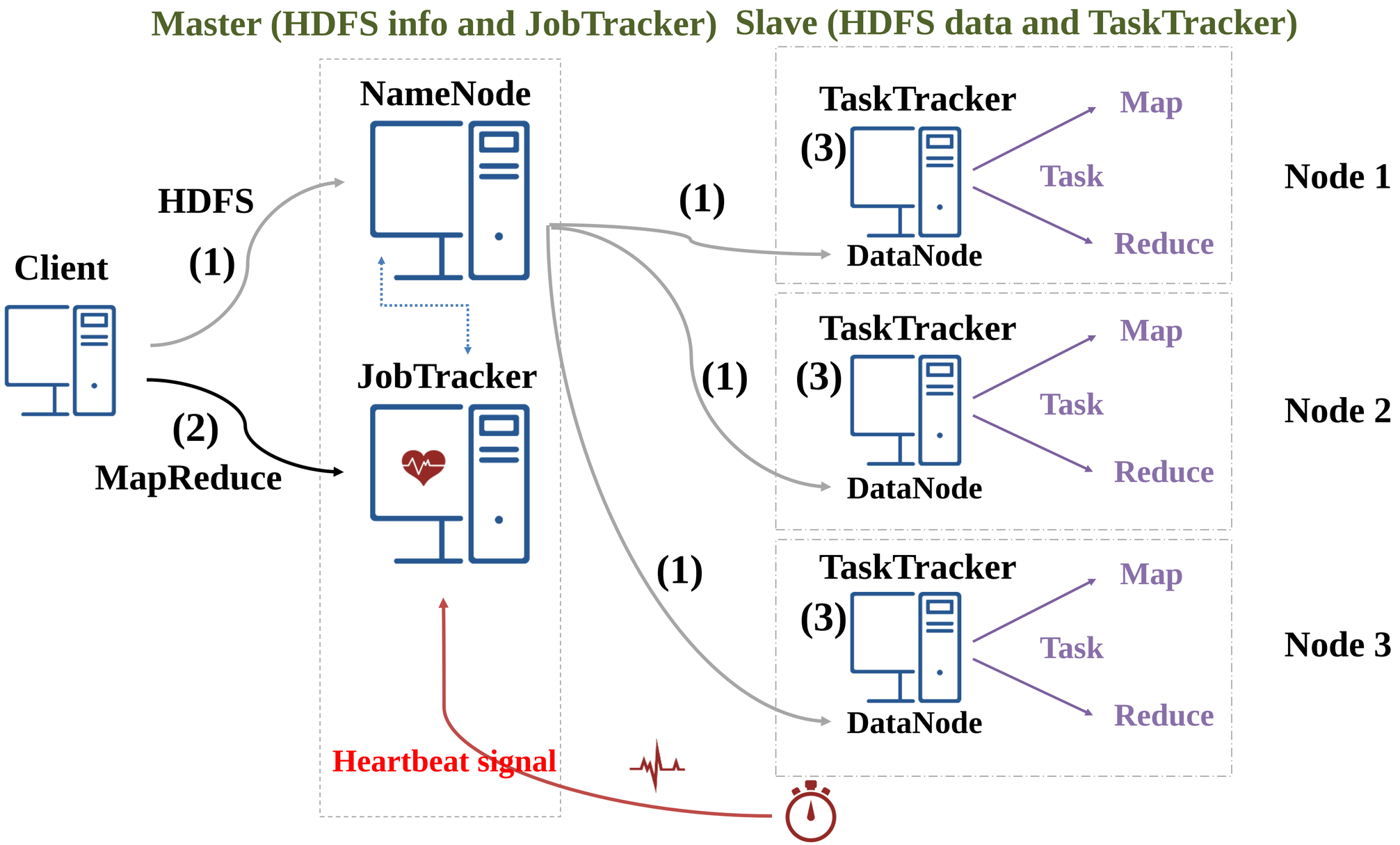
MapReduce components

- **Job Tracker**
  - Client applications submit jobs to JobTracker
  - JobTracker talks to NameNode to determine data location
  - JobTracker locates TaskTracker nodes with available slots at or near data.
  - JobTracker submits work to chosen TaskTracker nodes.
  - When work is completed, JobTracker updates status.

- **Task Tracker nodes are monitored:**
  - If they do not submit heartbeat signals periodically, they are deemed to have failed and work is scheduled on different TaskTracker

- **Task Tracker will notify JobTracker when a task fails:**
  - JobTracker may resubmit job elsewhere, mark that specific record as something to avoid, and even blacklist TaskTracker as unreliable.

Hadoop Scheme



JobTracker Workflow

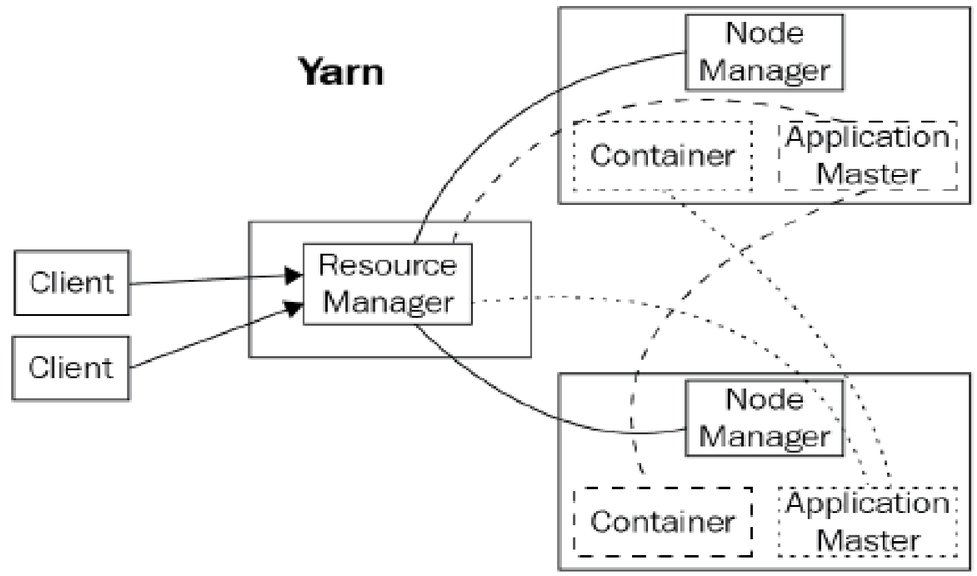
1. User applications submit jobs to the **JobTracker**
2. **JobTracker** talks to the **NameNode** to determine the location of the data
3. **JobTracker** locates **TaskTracker** with available slots in clusters
4. **JobTracker** submits the work to the chosen **TaskTracker** nodes
5. **TaskTracker** nodes are monitored. If they do not submit heartbeat signals often enough, they are deemed to have failed and the work is scheduled on a different **TaskTracker**
6. A **TaskTracker** will notify the **JobTracker** when a task fails. The **JobTracker** decides what to do then: it may resubmit the job elsewhere, it may mark that specific record as something to avoid, and it may even blacklist the **TaskTracker** as unreliable
7. When the work is completed, the **JobTracker** updates its status
8. The **JobTracker** is a point of failure for the Hadoop **MapReduce** service. If it goes down, all running jobs are halted

TaskTracker Workflow

1. A **TaskTracker** is a node in the cluster that accepts tasks - Map, Reduce and Shuffle operations from a **JobTracker**
2. When the **JobTracker** tries to find somewhere to schedule a task within the MapReduce operations, it first looks for an empty slot on the same server that hosts the **DataNode** containing the data, and if not, it looks for an empty slot on a machine in the same rack
3. The **TaskTracker** spawns a separate JVM processes to do the actual work; this is to ensure that process failure does not take down the task tracker
4. The **TaskTracker** monitors these spawned processes, capturing the output and exit codes. When the process finishes, successfully or not, the tracker notifies the **JobTracker**.
5. The **TaskTrackers** also send out heartbeat messages to the **JobTracker**, usually every few minutes, to reassure the **JobTracker** that it is still alive. These message also inform the **JobTracker** of the number of available slots, so the **JobTracker** can stay up to date with where in the cluster work can be delegated.

Apache Hadoop (2.0 & 3.0) YARN

- The fundamental idea of YARN (Yet Another Resource Negotiator) is to split up the functionalities of resource management and job scheduling/monitoring into separate daemons. It stands for **'next-generation MapReduce'.**



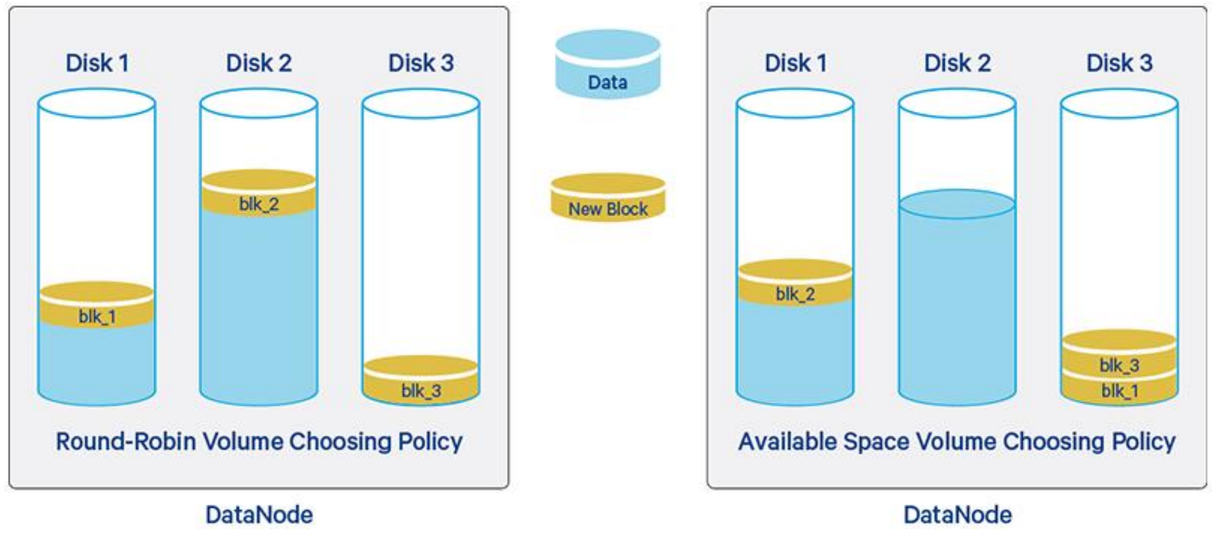
- **Resource Manager (RM)** knows where the DataNodes/Slaves are located (Rack Awareness) and how many resources they have. It decides how to assign the resources.
- **Node Manager (NM)** cares of the individual compute nodes, keeping up-to-date with the RM, monitoring resource usage (memory, CPU) in Container, tracking node-health by sending heartbeat signal to RM, log's management and auxiliary services.
- **Containers** is a fraction of computational resources (e.g., CPUs, memory, etc.)
- **Application Master** is responsible for the execution of a single application. It asks for containers and executes the application.

Hadoop 3.x against 1.x and 2.x

- **High Availability for NameNode:**
  - The loss of NameNode can crash the cluster in both Hadoop 1.x and 2.x.
  - Hadoop 2.x: the active-passive setup (one active and one stand-by NameNode) to help recover NameNode failures.
  - Hadoop 3.x: 1 active NameNode + 2 passive NameNodes + five JournalNodes to assist for **catastrophic failures**:
    - **NameNode machines**: run active and passive (standby) nodes. They should have the equivalent hardware to each other.
    - **JournalNode machines**: it is relatively lightweight and collects NameNode and YARN ResourceManager (JobTracker in the old version).

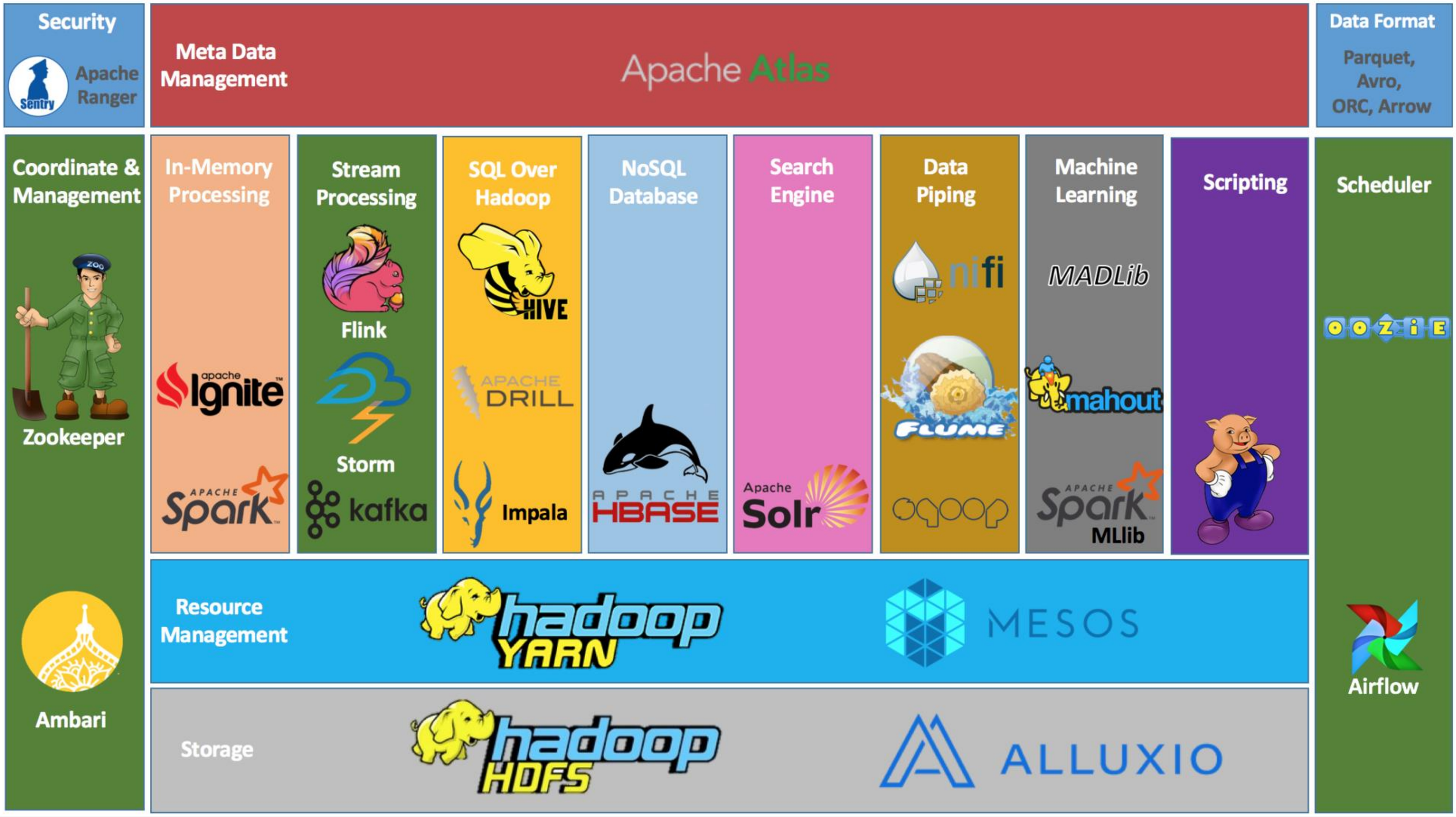
Intra-DataNode Balancer in Hadoop 3.x

- HDFS 3.x now includes a comprehensive storage capacity-management approach for moving data across nodes.
- When writing new blocks to HDFS, DataNode uses a volume-choosing policy to choose the disk for the block. Two such policy types are currently supported: *round-robin* or *available space*.



- The *round-robin* policy distributes the new blocks evenly across the available disks, while the *available-space* policy preferentially writes data to the disk that has the most free space by percentage.

Hadoop Ecosystem



Big Companies offer Big Data services

- Cloudera
- Databricks
- IBM
- Cloud Services: Amazon, Google, Microsoft

Hadoop Case Studies in Industry

BT

BT uses a Cloudera enterprise data hubpowered by Apache Hadoop to cut down on engineer call-outs.

By analysing the characteristics of its network, BT can identity whether slow internet speeds are caused by a network or customer issue. They can then evaluate whether an engineer would be likely to repair the problem.

The Cloudera hub provides a unified view of customer data stored in a Hadoop environment. BT earned a return on investment of between 200 and 250 percent within one year of the deployment.

BT has also used it to create new services such as "View My Engineer", an SMS and email alerting system that lets customers track the location of engineers. The company now wants to use predictive analytics to improve vehicle maintenance.

Source: computerworld.com

Hadoop Case Studies in Industry

Royal Bank of Scotland

The Royal Bank of Scotland (RBS) has been working with Silicon Valley company Trifacta to get its Hadoop data lake in order, so it can gain insight from the chat conversations its customers are having with the bank online.

RBS stores approximately 250,000 chat logs plus associated metadata per month. The bank stores this unstructured data in Hadoop. However, before turning to Trifacta this was a huge and untapped source of information about its user base

Source: computerworld.com

Hadoop Case Studies in Industry

CERN

The Large Hadron Collider in Switzerland is one of the largest and most powerful machines in the world. It is equipped with around 150 million sensors, producing a petabyte of data every second, and the data being delivered is growing all the time.

CERN researcher Manuel Martin Marquez said: “This data has been scaling in terms of amount and complexity, and the role we have is to serve to these scaleable requirements, so we run a Hadoop cluster.”

“From a simplistic manner we run particles through machines and make them collide, and then we store and analyse that data.”

“By using Hadoop we limit the cost in hardware and complexity in maintenance.”

Source: computerworld.com

Hadoop Case Studies in Industry

Royal Mail

British postal service company Royal Mail has used Hadoop to get the "building blocks in place" for its big data strategy.

Director of the Technology Data Group at Royal Mail, Thomas Lee-Warren, told Computerworld UK that its Hadoop investment is the foundation of a drive to gain more value from internal data. "We have a lot of data," Lee-Warren explained. "We are about to go up to running in the region of a hundred terabytes, across nine nodes."

The business uses Hortonworks' Hadoop analytics tools to transform the way it manages data across the organisation, freeing the analytics team to deliver insights on proprietary information held in its data warehouse.

Source: computerworld.com

Hadoop Case Studies in Industry

British Airways

British Airways deployed its first instance of Hadoop in April 2015, as a data archive for legal cases that were primarily stored, at a high cost, on its enterprise data warehouse (EDW) platform.

Since deploying Hortonworks 2.2 HDP, Spanos said his department has returned on its investment within a year, and is able to deliver 75 percent more free space for new projects, which translates to cost reductions to the airline's finance team.

British Airways' data exploitation manager Alan Spanos said: "In business intelligence, if you don't adopt this technology to do at least part of your job role, you will not exist in a few years' time. You can only go so far with traditional technology. It still has a place within your architecture, but quite frankly, this is where you need to be."

Source: computerworld.com

Hadoop Case Studies in Industry

Western Union

Global payments provider Western Union implemented a Hadoop-based data analytics platform from Cloudera in 2014 to provide a more personalised experience for its customers.

Using Cloudera Enterprise, Western Union is able to more efficiently store and process real-time analytics on what the vendor describes as “one of the world’s largest enterprise data sets”.

Cloudera’s Apache Hadoop implementation helps Western Union centralise its global customer data in an enterprise data hub, and supports pattern recognition and predictive modelling. The big data analytics platform is aimed at creating a more personalised experience across multiple products and service delivery channels for Western Union customers.

Source: computerworld.com

Hadoop Case Studies in Industry

King.com

European gaming giant and creator of Candy Crush King.com deployed Cloudera's Distribution for Apache Hadoop in 2012. The aim was to run analytics for every 'event', or action, its millions of users take during gameplay.

The company's director of data warehousing, Mats-Mats Eriksson, told Computerworld UK that using analytics is vital to its success online.

"Analytics is one of the things that made King.com the thing that it is today," Eriksson explained. "In the universe that we operate in, gaming online, it is absolutely essential to know as much as possible about the players and optimise everything."

"Everybody wants a business case for Hadoop, but for me it is simply about difference between knowing what happens in a game and not knowing."

Source: computerworld.com

Hadoop Case Studies in Industry

Expedia

Expedia planned to double its Hadoop investment back in 2015and was an early adopter of Hortoworks project Apache Falcon to crunch large volumes of numbers.

Expedia previously used a DB2 database in conjunction with various instances of Microsoft SQL server, which became increasingly expensive to scale as data volume increased with the business growing organically, along with acquiring several travel companies including Trivago and Hotels.com.

Since moving to Hadoop, the firm has seen costs drop and is able to both store and process data using the cluster.

Woodhead, who is data platform technical lead for Hotels.com, revealed that "hundreds" of employees across different departments and offices, one of which is based in London, used the two-petabyte cluster for web traffic, bookings and travel reviews.

Source: computerworld.com

Hadoop Case Studies in Industry

Hotels.com

Hotels.com uses Hadoop for huge data storage and offline analytics - that means crunching large amounts of data and not expecting an answer within a millisecond. Cassandra, on the other hand, is used in the online transactional world “where you need an answer below ten milliseconds”.

It can also store the data, but is targeted at online for its speedy capabilities. The business moved from traditional relational databases like Microsoft SQL server three years ago to become “active/active”.

Chief technology officer at Hotels.com, Thierry Bedos, said: "We started solving a real issue for the business - which was customer service and personalising what we offer them online - whereas some firms use big data as an innovation project and say 'we need to play with big data, let's think of some cool use cases we think will add value'".

Source: computerworld.com

Hadoop Case Studies in Industry

Marks and Spencer

Retail giant M&S adopted the Cloudera Enterprise Data Hub Edition system in 2015 to analyse data from multiple sources, to better understand customer behaviour.

Jagpal Jheeta, head of business information and customer insight at M&S, said: “Smart and efficient data usage is a key focus at M&S, as it ultimately fuels better customer insight, engagement and loyalty. We needed a scalable, robust and future-proof strategic partner. Cloudera is aiding us in leveraging analytics to better serve the business now and in the future.”

Source: computerworld.com

Hadoop Case Studies in Industry

Tesla

Tesla is using a Hadoop clusterto collect the increasing amount of data being generated by its connected cars.

CIO Jay Vijayan said: “We are working on a big data platform... The car is connected, but it does not really talk to the network every minute because we want to keep it as smart and efficient as possible. It alerts us if the car is not functioning properly so service teams can take action.”

Source: computerworld.com

Summary

- Discussed the reasons for Big Data technologies
- Introduced Hadoop
- Discussed Hadoop Infrastructure
- Discussed HDFS and MapReduce
- Introduced Hadoop Ecosystem Components