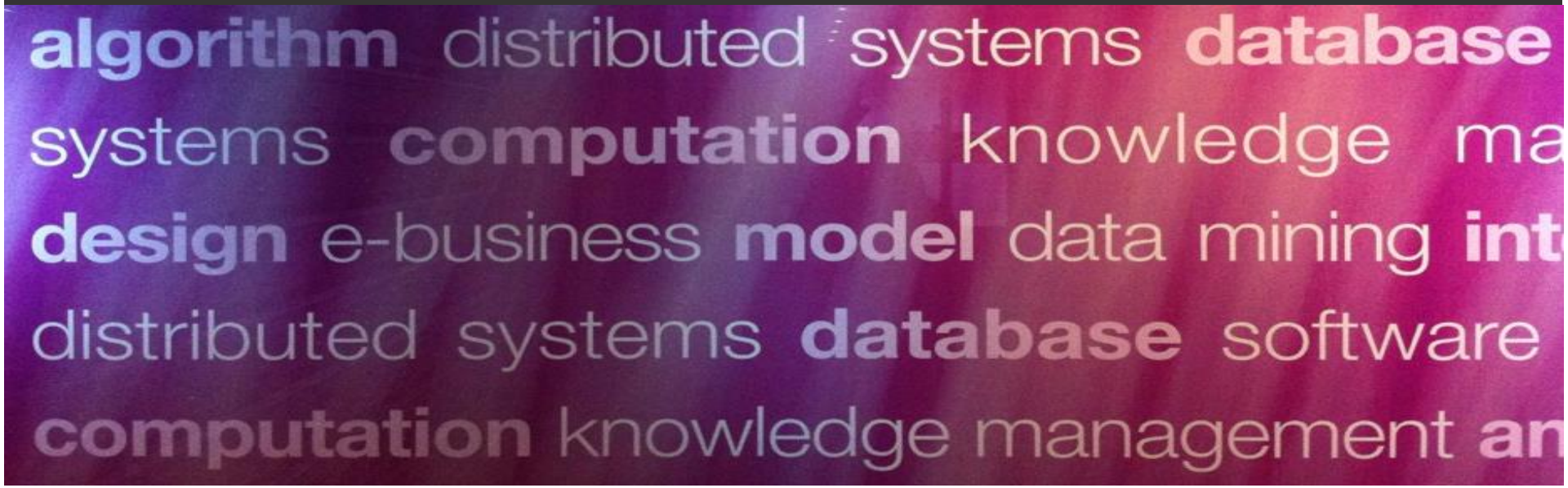
MONASH University

Information Technology

FIT5202

Week 1b – Introduction to Big Data



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

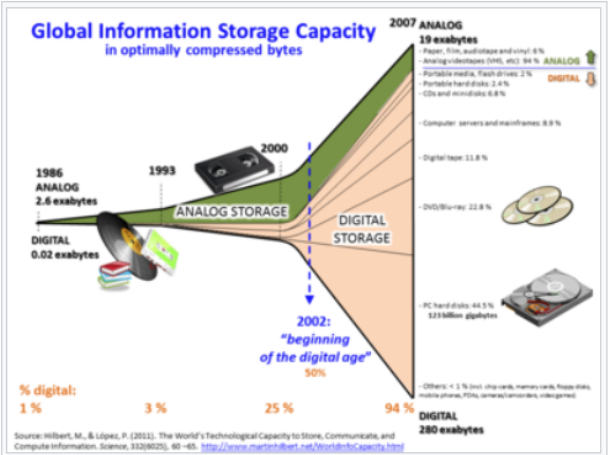
Print/export
Create a book
Download as PDF
Printable version

Big data

From Wikipedia, the free encyclopedia

This article is about large collections of data. For the band, see Big Data (band).

Big data is data sets that are so voluminous and complex that traditional data processing application software are inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy. There are three dimensions to big data known as Volume, Variety and Velocity.



Growth of and digitization of global information-storage capacity[1]

Lately, the term "big data" tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem."[2] Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on."[3] Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data-sets in areas including Internet search, fintech, urban informatics, and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics,[4] connectomics, complex physics simulations, biology and environmental research.[5]

Data sets grow rapidly - in part because they are increasingly gathered by cheap and numerous information-sensing Internet of things devices such as mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks.[6][7] The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s;[8] as of 2012, every day 2.5 exabytes (2.5×10^{18}) of data are generated.[9] By 2025, IDC predicts there will be 163 zettabytes of data.[10] One question for large enterprises is determining who should own big-data

This unit is about… (Main learning outcomes)

1. **Volume** ➔ Weeks 1, 2, 3, 4
   - How to process Big Data Volume?

Assignment 1 (15 %)

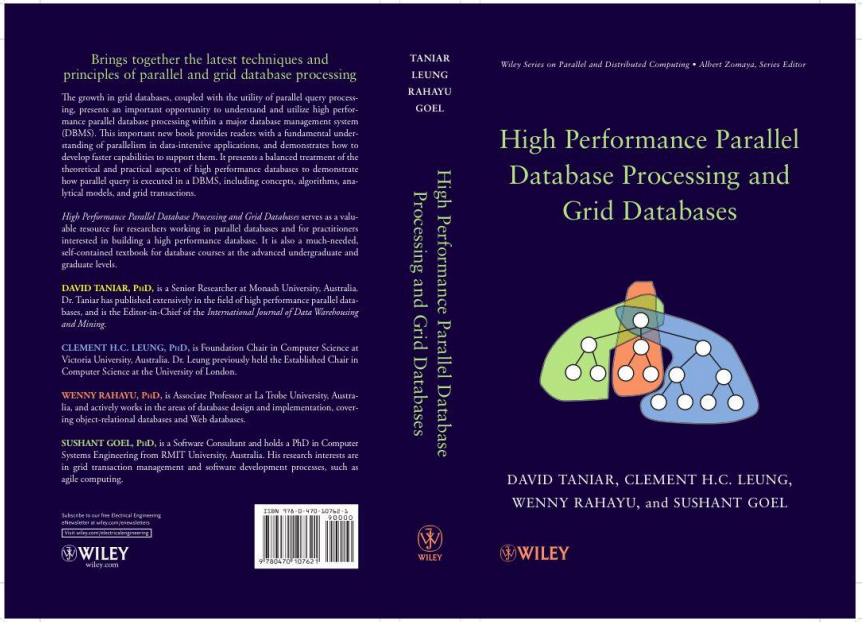
2. **Complexity** ➔ Weeks 5, 6, 7, 8
   - How to apply machine learning algorithms to every aspect of Big Data?

Assignment 2 (30 %)

3. **Velocity** ➔ Weeks 9, 10, 11
   - How to handle and process Fast Streaming Data?

This unit is about…

1. **Volume** ➔ Weeks 1, 2, 3, 4
   - How to process Big Data Volume
   - Parallel Algorithms
   - **Textbook**: https://onlinelibrary-wiley-com.ezproxy.lib.monash.edu.au/doi/book/10.1002/9780470391365



What is Big Data Volume?

"**Everyday**, 2.5 **quintillion** bytes of data are created and 90% of the data in the world today was created within the past two years".

IBM Corporation

10^6 = million (megabytes)
10^9 = billion (gigabytes)
10^{12} = trillion (terabytes)
10^{15} = quadrillion (petabytes)
10^{18} = **quintillion** (exabytes)

What is Big Data Volume?

"**Everyday**, 2.5 **quintillion** bytes of data are created and 90% of the data in the world today was created within the past two years".

IBM Corporation

"Worldwide information is more than **doubling every two years**, with **4.4 zettabytes** in 2013 to 44 zettabytes by 2020"; More data will be created in 2017 than the previous 5,000 years of humanity.

Developer Magazine

…
10^{15} = quadrillion (petabytes)
10^{18} = **quintillion** (exabytes)
10^{21} = sextillion (**zettabytes**)

What is Big Data Volume?

Data comes from everywhere:

- Post to **social media** sites

"As of April 2020, Facebook tops **2.89 billion** active monthly users"

exalteddigital

"Twitter has over 330 million monthly active users in 2020, generating over **500 million tweets** and handling over 2.1 billion search queries per day".

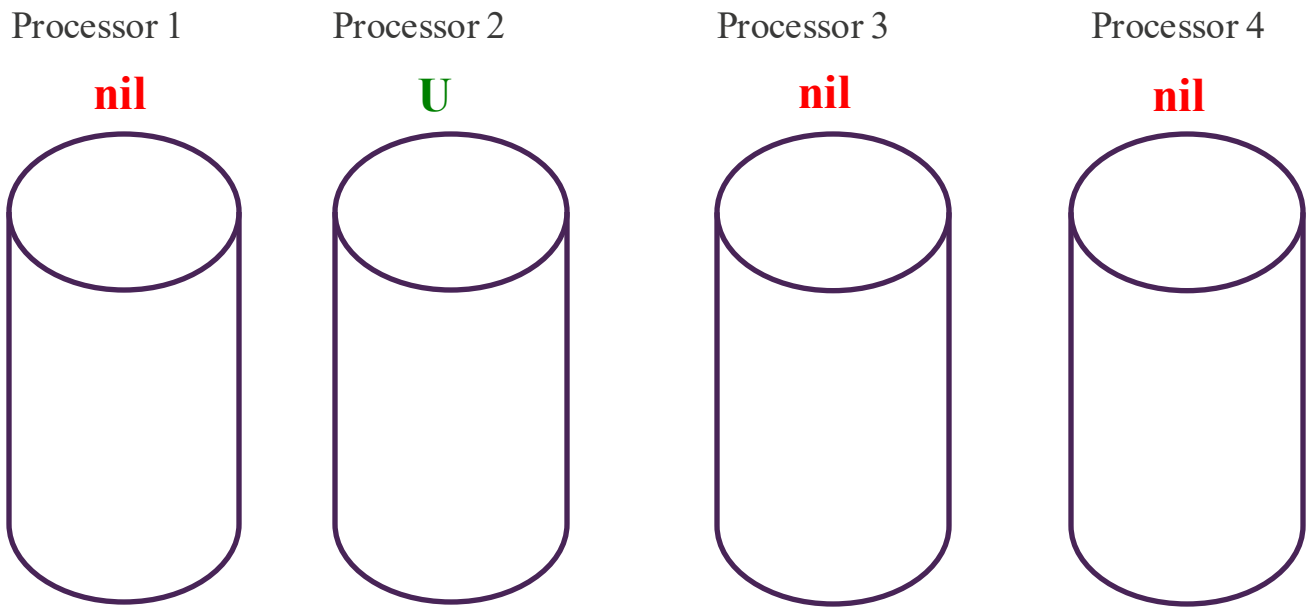
Twitter wikipedia

How to process Big Data Volume?

- **Parallel Databases**
- Parallelization through data partitioning
- Hence, parallel scans, yield **I/O parallelism**

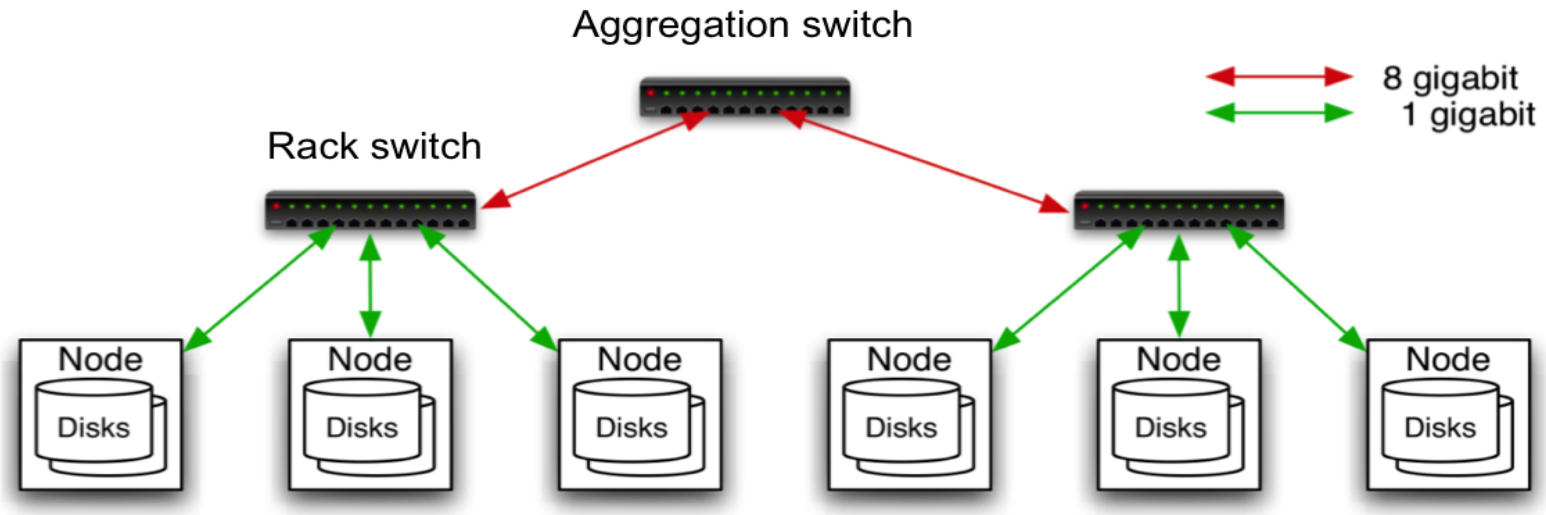
1. Data partitioning
2. Processing data partitions in parallel

Search U



How to process Big Data Volume?

- Parallel computing
  - Constructing high performance parallel computers using a large number of (low-end) **commodity processors**.
  - Commodity machines (cheap, but unreliable).
  - Commodity network.
  - Scalable (1000's of machines, 10,000's of disks)



How to process Big Data Volume?

- Parallel programming
  - Parallel/Distributed Programming in the past: MPI
  - A new parallel programming paradigm: **MapReduce**

- MapReduce: a simple data-parallel programming model designed for **scalability** and **fault-tolerance**.

- Pioneered by Google
  - Processes 20 Petabytes of data per day
- Popularized by open-source Hadoop project
  - Used at Yahoo!, Facebook, Amazon, …

MapReduce

- **Cheap nodes fail**, especially if you have many of them
  - Mean time between failures for 1 node = 3 years
  - Mean time between failures for 1000 nodes = 1 day
  - ***Solution***: Build fault-tolerance into system

- **Commodity network = low bandwidth**
  - ***Solution***: Push computation to the data (localized computation to avoid much transfer between machines)

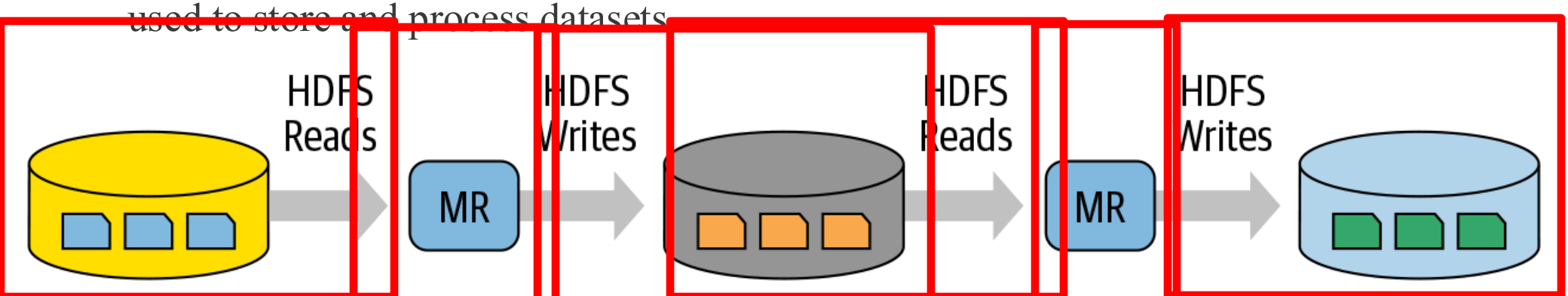
- **Programming distributed systems is hard**
  - ***Solution***: Data-parallel programming model users write "map" and "reduce" functions, system distributes work and handles faults.

MapReduce and Apache Hadoop

- **Map Reduce** is a programming model for large scale parallel processing of Data. The model consist of two functions Map and Reduce. Mapper is a function that performs filtering and Reducer groups the data provided by Mapper.

- **Hadoop** is an open source implementation of Map Reduce. Map Reduce is one of the core components of Hadoop system.

- The other core component is **Hadoop Distributed File System** (**HDFS**), used to store and process datasets.



Apache Spark

- **Apache Spark** is a Big Data distributed processing framework that supports reuse of working set of data across multiple parallel operations.
- It supports
  - Batch processing (Spark Core)
  - Real-time stream processing (Spark Streaming)

Metrics	Apache Hadoop	Apache Spark
Speed		✓
Ease of Use		✓
Generality		✓
Runs Everywhere		✓
Scheduler	✓	✓
API	✓	✓
Fault Tolerance	✓	✓
Maturity	✓	

Hadoop vs. Spark

Figure 4. Performance of logistic regression in Hadoop MapReduce vs. Spark for 100GB of data on 50 `m2.4xlarge` EC2 nodes.

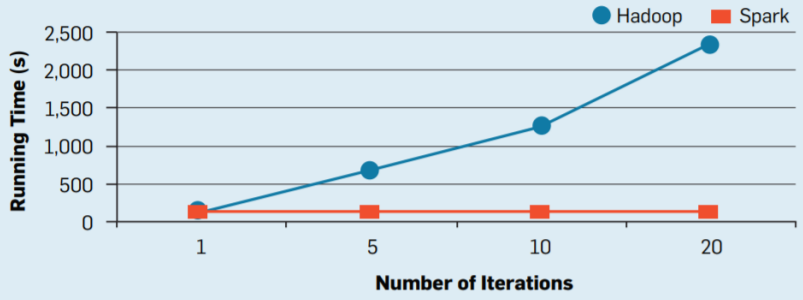
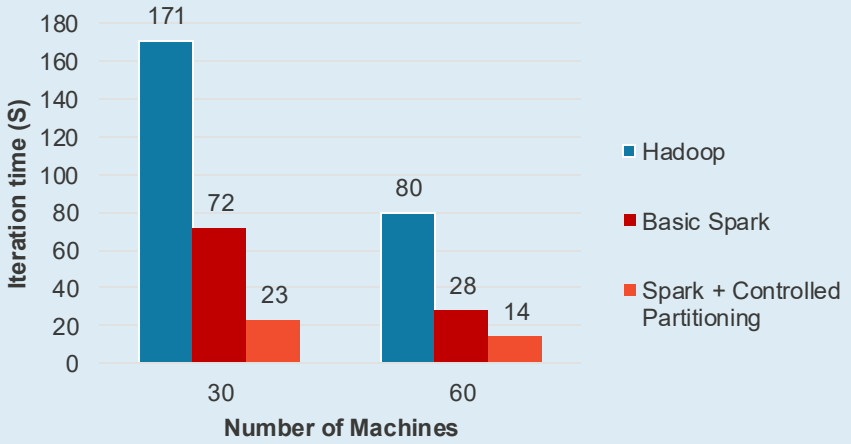


Figure 10: Performance of PageRank on Hadoop and Spark



[1] Zaharia, Matei, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng et al. "Apache Spark: A unified engine for big data processing." *Communications of the ACM* 59, no. 11 (2016): 56-65.

[2] Zaharia, Matei, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. Franklin, S. Shenker, and I. Stoica. "Resilient distributed datasets." In *A fault-tolerant abstraction for in-memory cluster computing in Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. 2014.

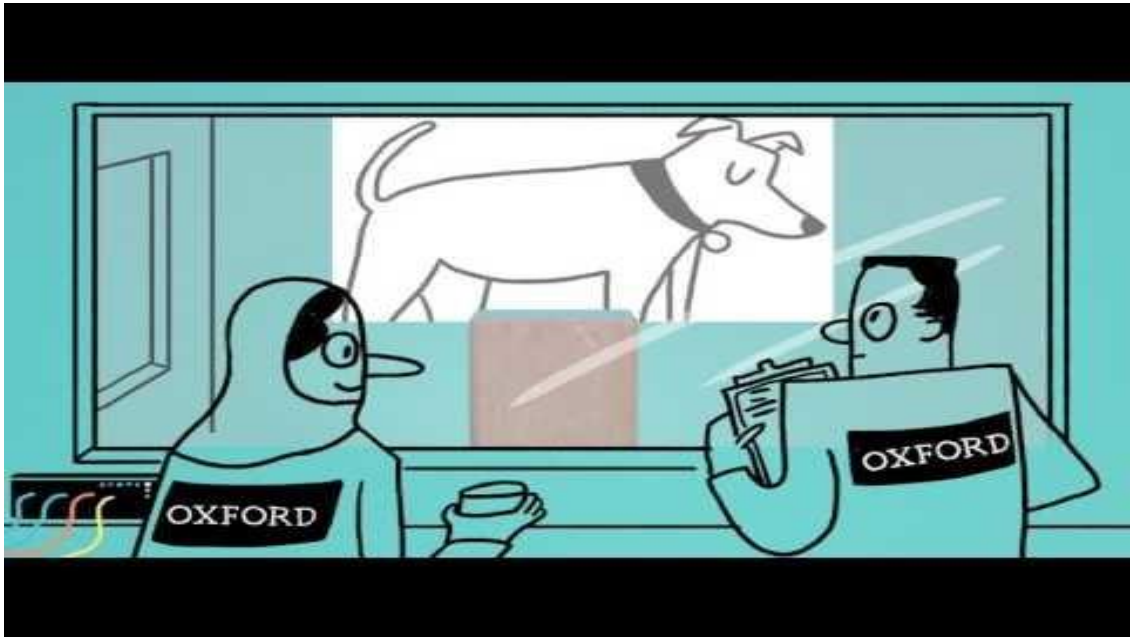
This unit is about…

1. **Volume** ➔ Weeks 1, 2, 3, 4
   - How to process Big Data Volume?

2. **Complexity** ➔ Weeks 5, 6, 7, 8
   - How to apply machine learning algorithms to every aspect of Big Data?

Machine Learning

- Machine learning algorithms attempt to make predictions or decisions based on *training data*, often maximizing a mathematical objective about how the algorithm should behave.
- There are multiple types of learning problems:
  - **Classification**
  - **Regression**
  - **Clustering** etc.



Machine Learning Pipeline

An example of classification: Whether an email is spam or non-spam based on labeled examples of other items (e.g., emails known to be spam or not).

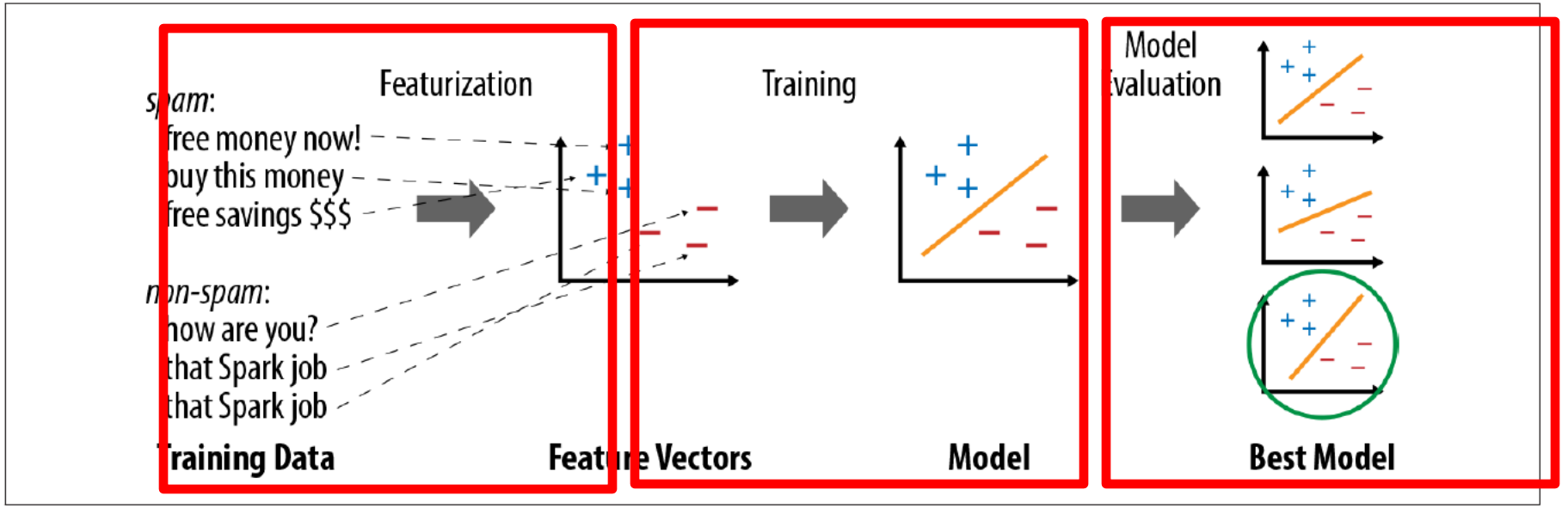


Figure 11-1. Typical steps in a machine learning pipeline

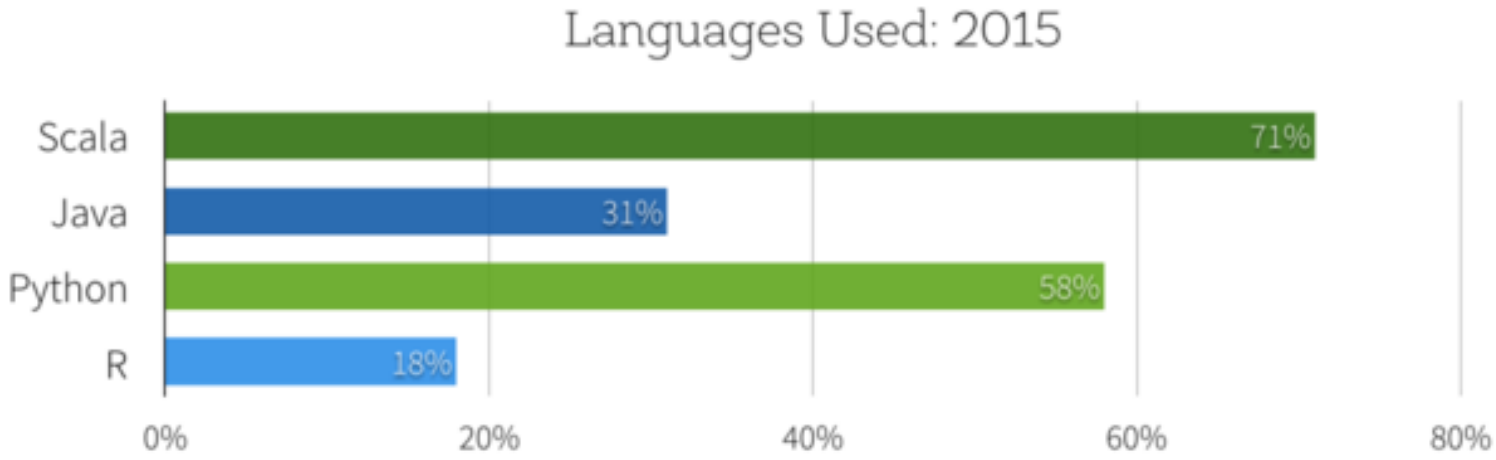
Spark for Machine Learning?

- The traditional uses of Python or R tools are often limiting.
- They process data on a single machine where the
  - **movement of data becomes time consuming**,
  - **the analysis requires sampling** and
  - moving from development to production environments requires extensive **re-engineering**.

- Spark MLlib enhances machine learning because of its **simplicity**, **scalability**, and **easy integration** with other tools.

Spark for Machine Learning?

- Spark also provides many language choices, including Scala, Java, Python, and R.



Source: 2015 Spark Survey

- Apache **Spark** is a unified analytics engine for large-scale data processing. It provides high-level **APIs** in Java, Scala, Python and R

This unit is about…

1. **Volume** ➔ Weeks 1, 2, 3, 4
   - How to process Big Data Volume?

2. **Complexity** ➔ Weeks 5, 6, 7, 8
   - How to apply machine learning algorithms to every aspect of Big Data?

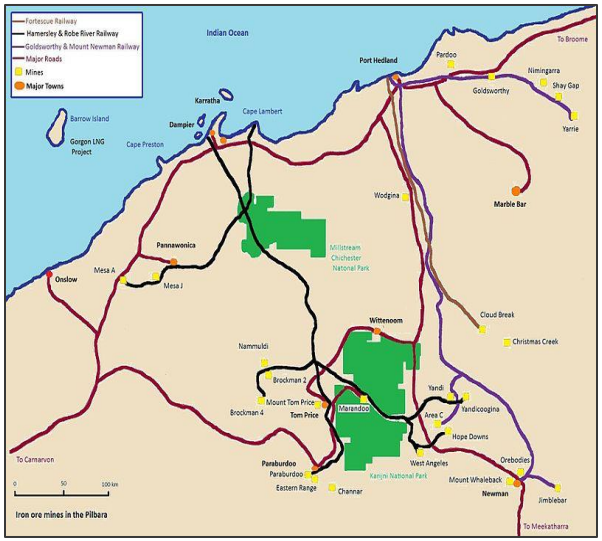
3. **Velocity** ➔ Weeks 9, 10, 11
   - How to handle and process Fast Streaming Data?

New Data Producers…

1. High speed data producers
   - Sensors
2. Characteristics
   - High speed data
   - High inaccuracy
   - Needs some pre-processing
3. Processing requirements
   - How to filter data
   - How to pre-process data
   - How to store data

More realistic projects…

Heavy-Haul Railway Project

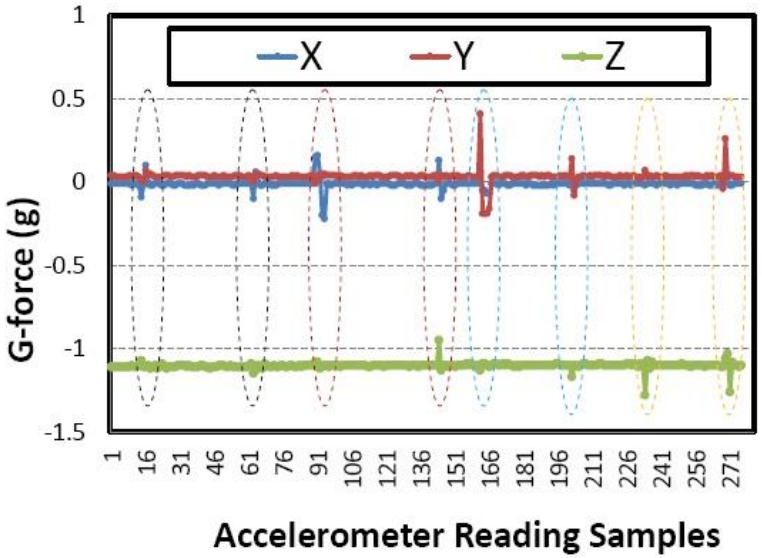


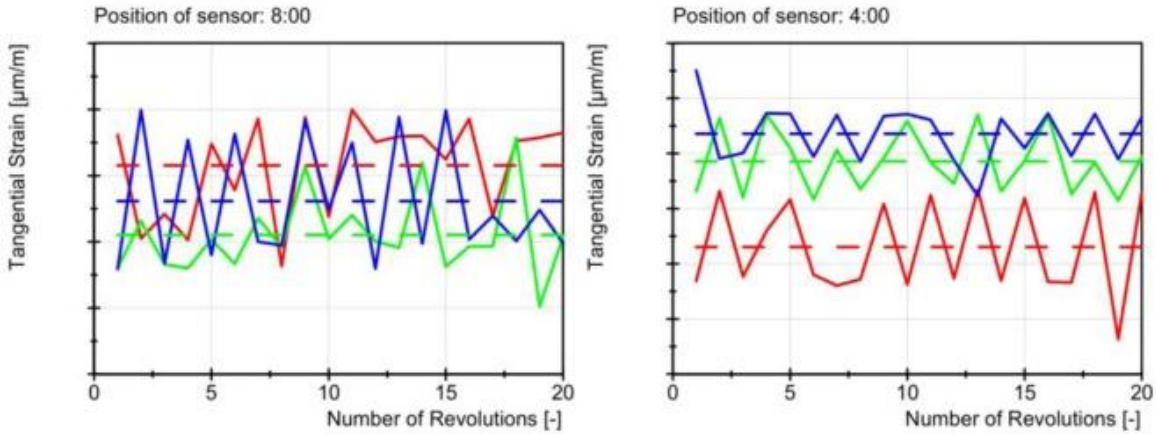
- Each car has 20-30 low-cost sensors, measuring acceleration, temperature, etc.
- The data is mostly static

More realistic projects…

Heavy-Haul Railway Project

- Sensor readings

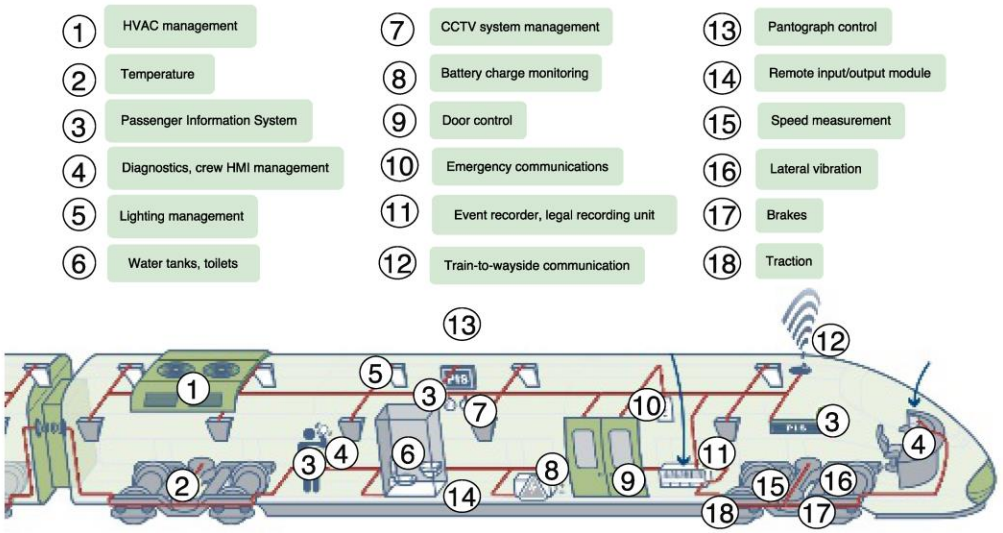


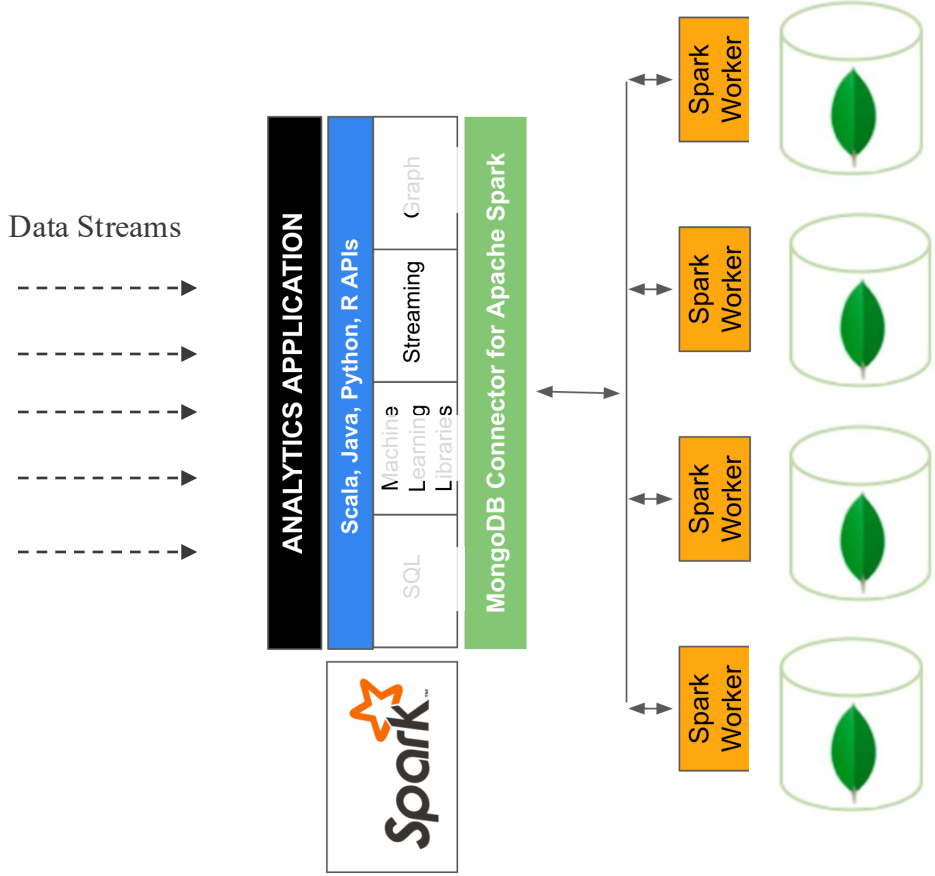


More realistic projects…

Challenges:
How to absorb the data quickly?
How to filter and pre-process data?
How to store data?

Heavy-Haul Railway Project





Summary of Big Data

1. **Volume**
   - Use Apache Sparks' parallel programming paradigm to process large volume of data
   - Use **Python** as the programming language

2. **Complexity**
   - Use **Spark MLlib** to learn from your Big Data

3. **Velocity**
   - Focus on Stream Data processing
   - Use Apache Kafka and **Spark Streaming** to handle the velocity of data