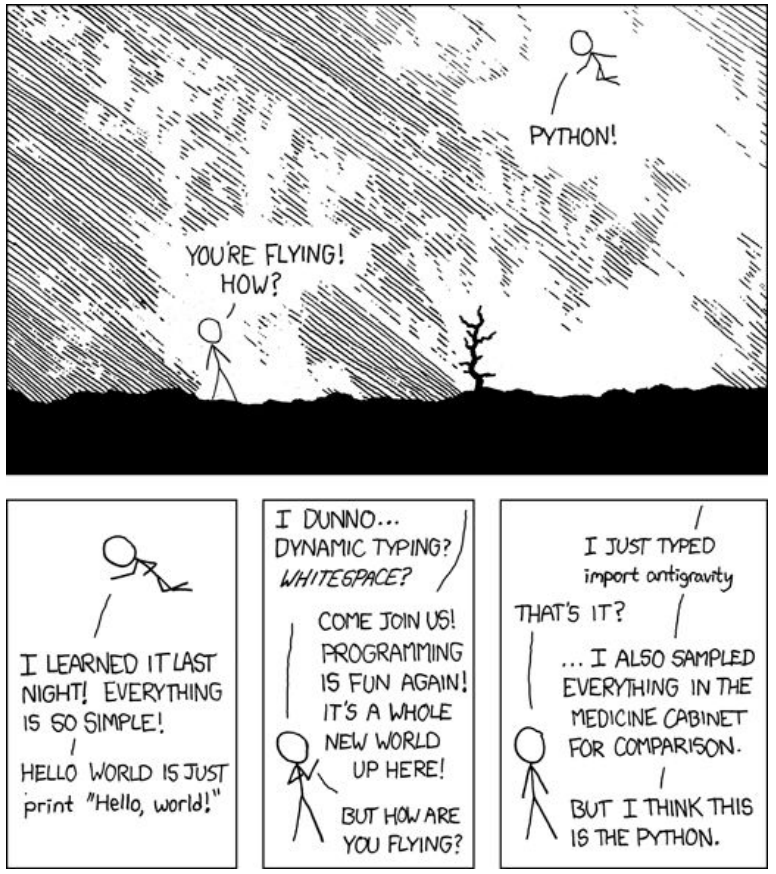
UMM
UNIVERSITÄTSMEDIZIN
MANNHEIM

Medizinische Fakultät Mannheim
der Universität Heidelberg

Universitätsklinikum Mannheim

Python for Big Data Analysis



Kim Hee
Graduate research assistant
at Heinrich-Lanz-Center (HLZ) for Digital Health

Tutorial organization

Time	Topic
09:00 - 09:30	Set the scene
09:30 - 10:45	Data analysis and Data visualization
10:45 - 11:15	Break
11:15 - 12:15	Machine Learning
12:15 - 13:00	**Python at Scale (PySpark)**

Agenda

- PySpark in Hadoop ecosystem
- Three core components of Hadoop cluster
- Demo and Hands-on

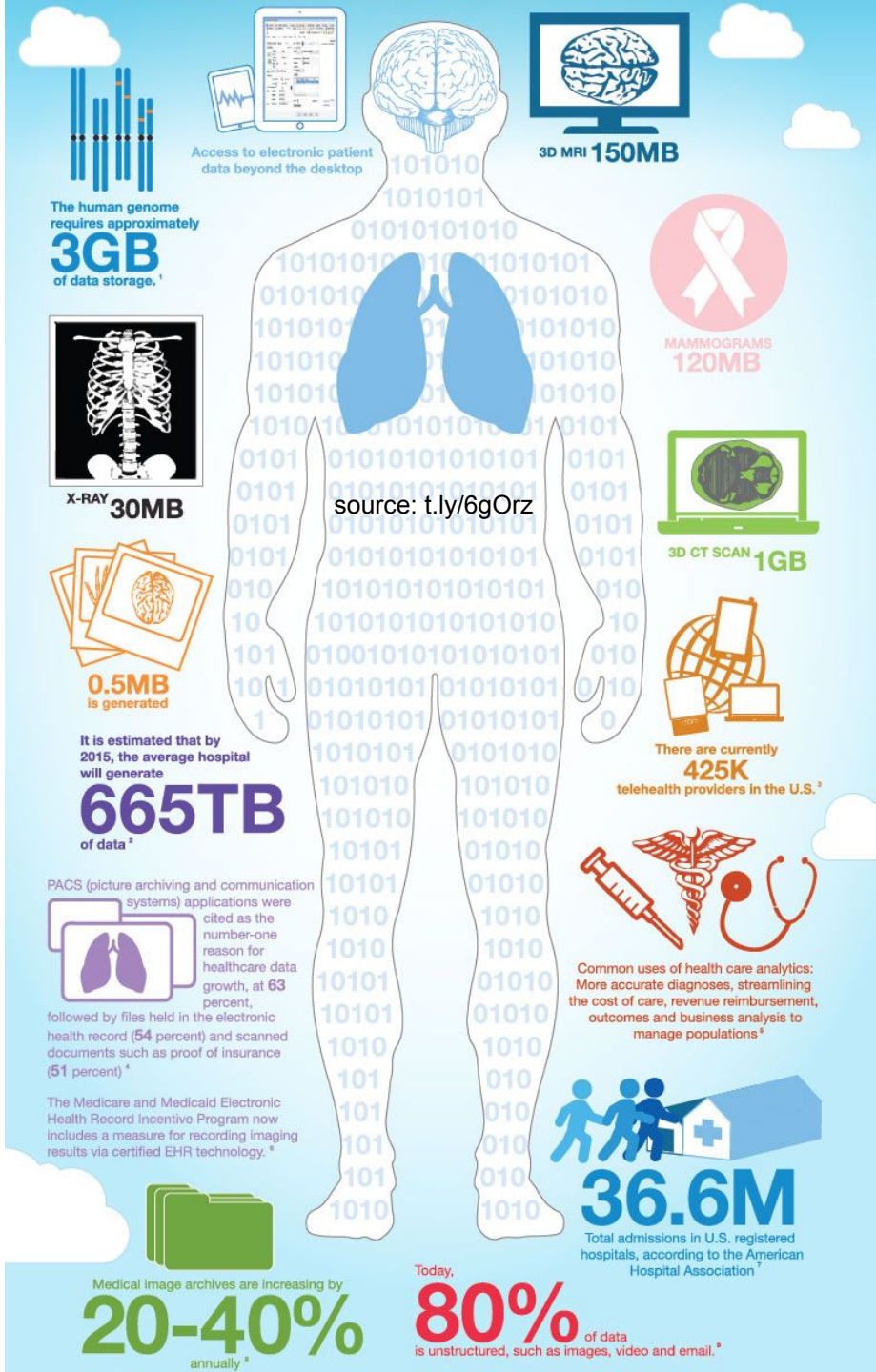
PySpark in Hadoop ecosystem
recall the modops movement

- $1 million solution has never been used. The solution improved the recommendation algorithm by 10%, but Netflix never implemented that solution.

> *“We evaluated some of the new methods offline but* ***the additional accuracy gains that we measured did not seem to justify the engineering effort*** *needed to bring them into a production environment.”*

PySpark in Hadoop ecosystem
the explosion of healthcare data

- Healthcare data is generated **annually 48% more** according to Stanford Medicine 2017 Health Trends Report
- 153 exabytes were produced in 2013 and an estimated **2,314 exabytes will be produced in 2020**
- **325,000 mobile health apps** available **in 2017** (t.ly/VrxrZ)
- Global medical app downloads exceeded **400M in 2018** (t.ly/Pbjb8)
- Genome sequence ≈ 3GB, a mammogram ≈ 120MB, a 3D MRI ≈ 150MB, a 3D CT-Scan ≈ 1GB

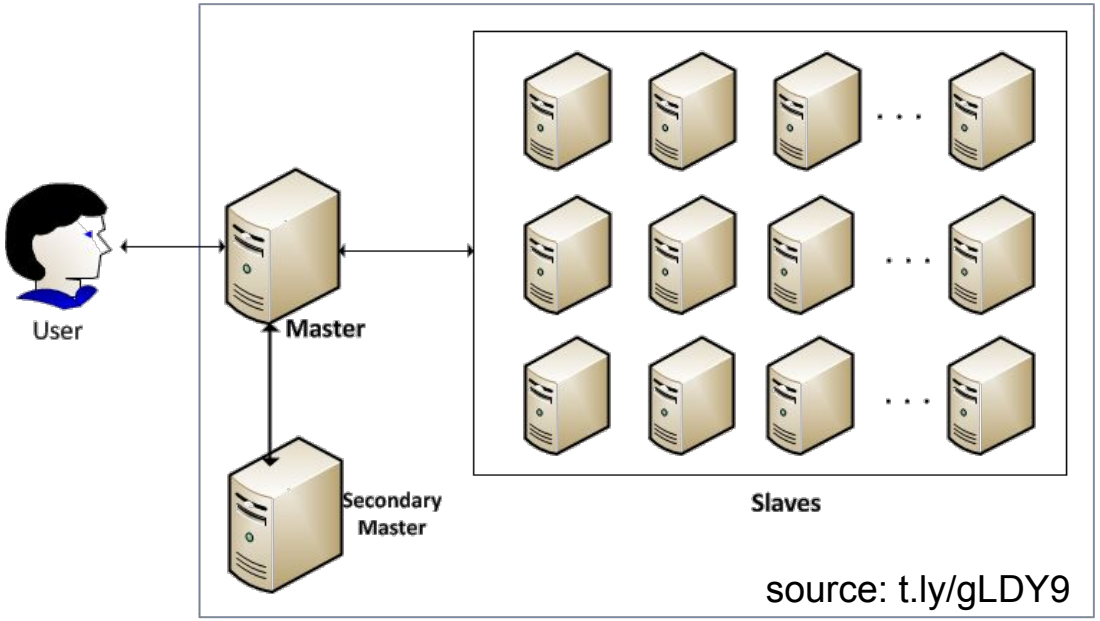


UMM
UNIVERSITÄTSMEDIZIN MANNHEIM

PySpark in Hadoop ecosystem
the distributed system beats the bigger computer

"In pioneer days they used oxen for heavy pulling, and when one ox couldn't budge a log, they didn't try to grow a larger ox. We shouldn't be trying for bigger computers, but for **more systems of computers.**"

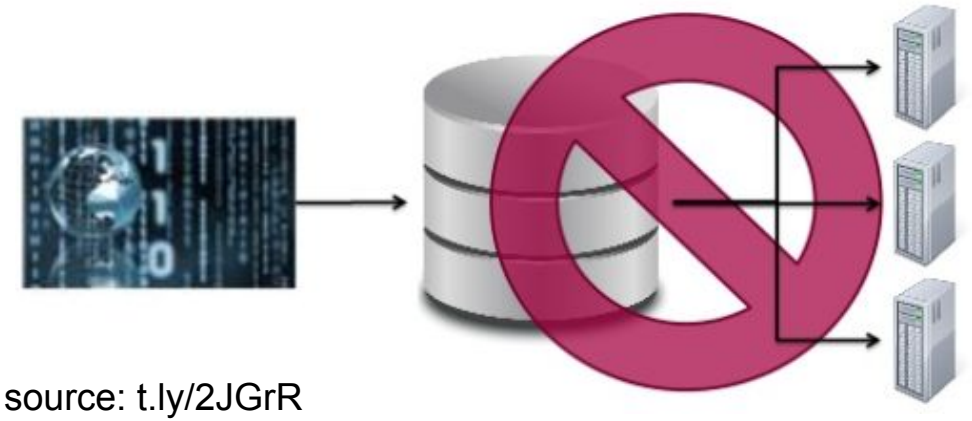
- Grace Hopper



source: t.ly/gLDY9

PySpark in Hadoop ecosystem
the bottleneck of a conventional distributed system

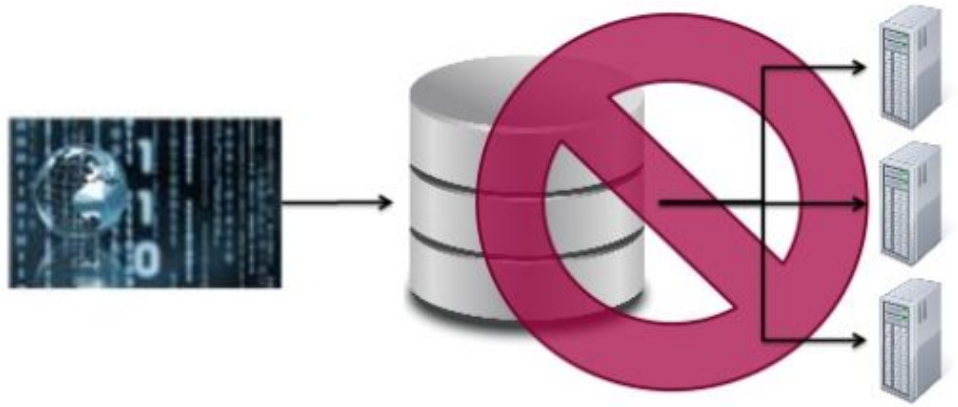
- Data are stored in a central location and copied to processors at runtime
- It is impossible to process terabytes or petabytes data due to the finite bandwidth



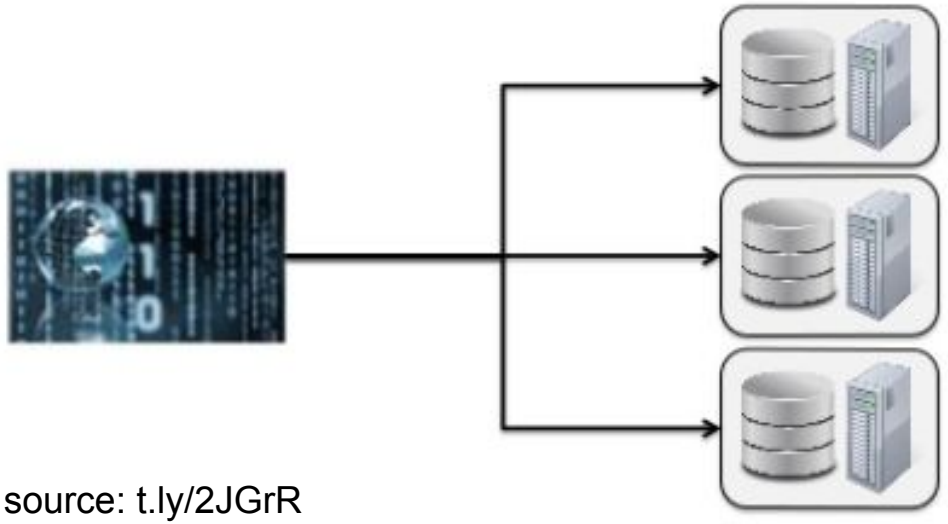
source: t.ly/2JGrR

PySpark in Hadoop ecosystem
the bottleneck of a conventional distributed system

- Data are stored in a central location and copied to processors at runtime
- It is impossible to process terabytes or petabytes data due to the finite bandwidth



- Hadoop introduced a radical approach that **brings the program to the data**
- It's distributed, scalable, fault-tolerant, and even open source
- Three core components of Hadoop cluster:
  - Storage; Processing; Resource management

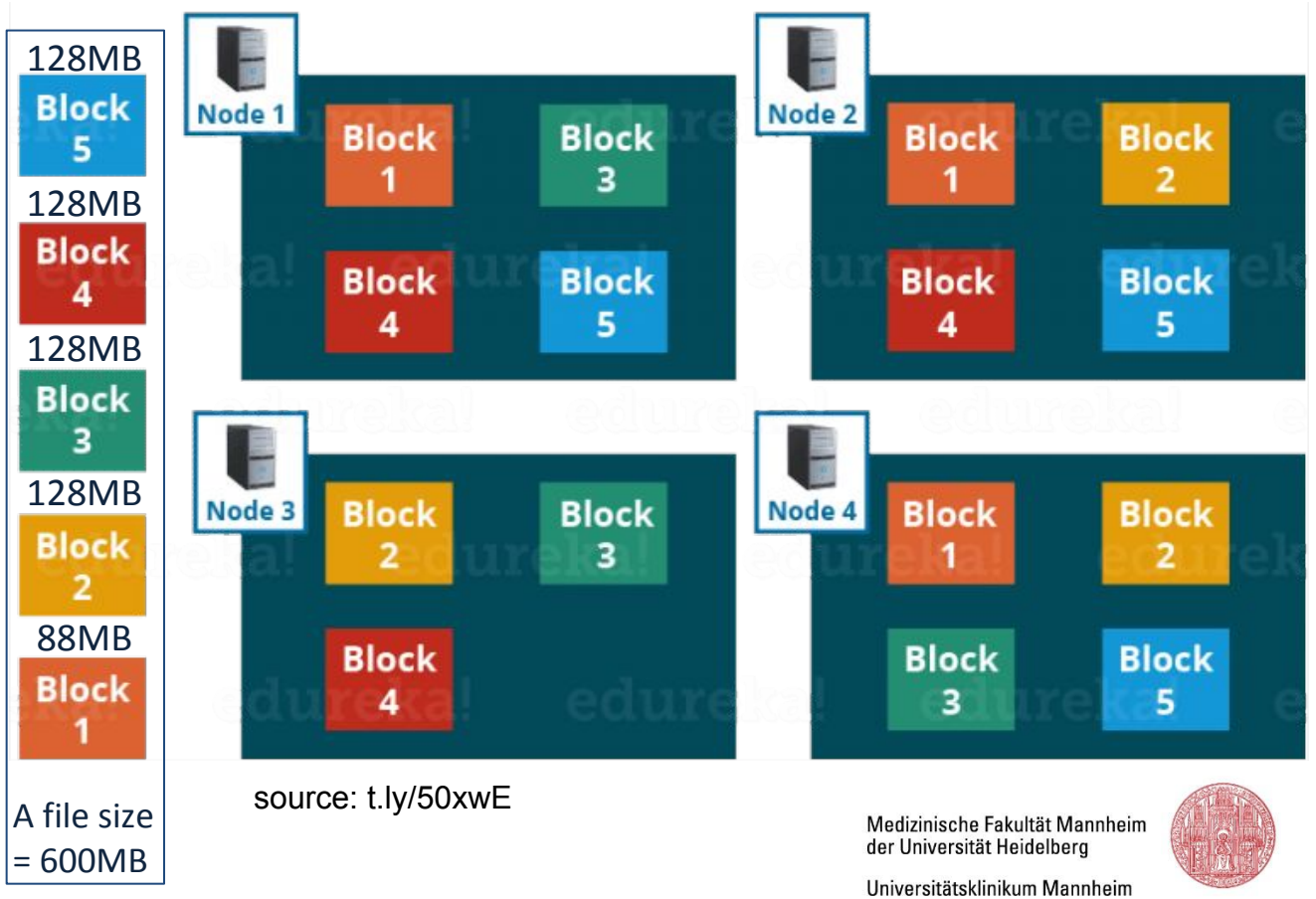


source: t.ly/2JGrR

Three core components of Hadoop cluster: **storage**, processing, and resource management

Hadoop Distributed File System (HDFS)

- HDFS is a filesystem based on Google's GFS
- Data is distributed when stored
- Data files are split into **128MB blocks**
- Each block is replicated (default 3x)



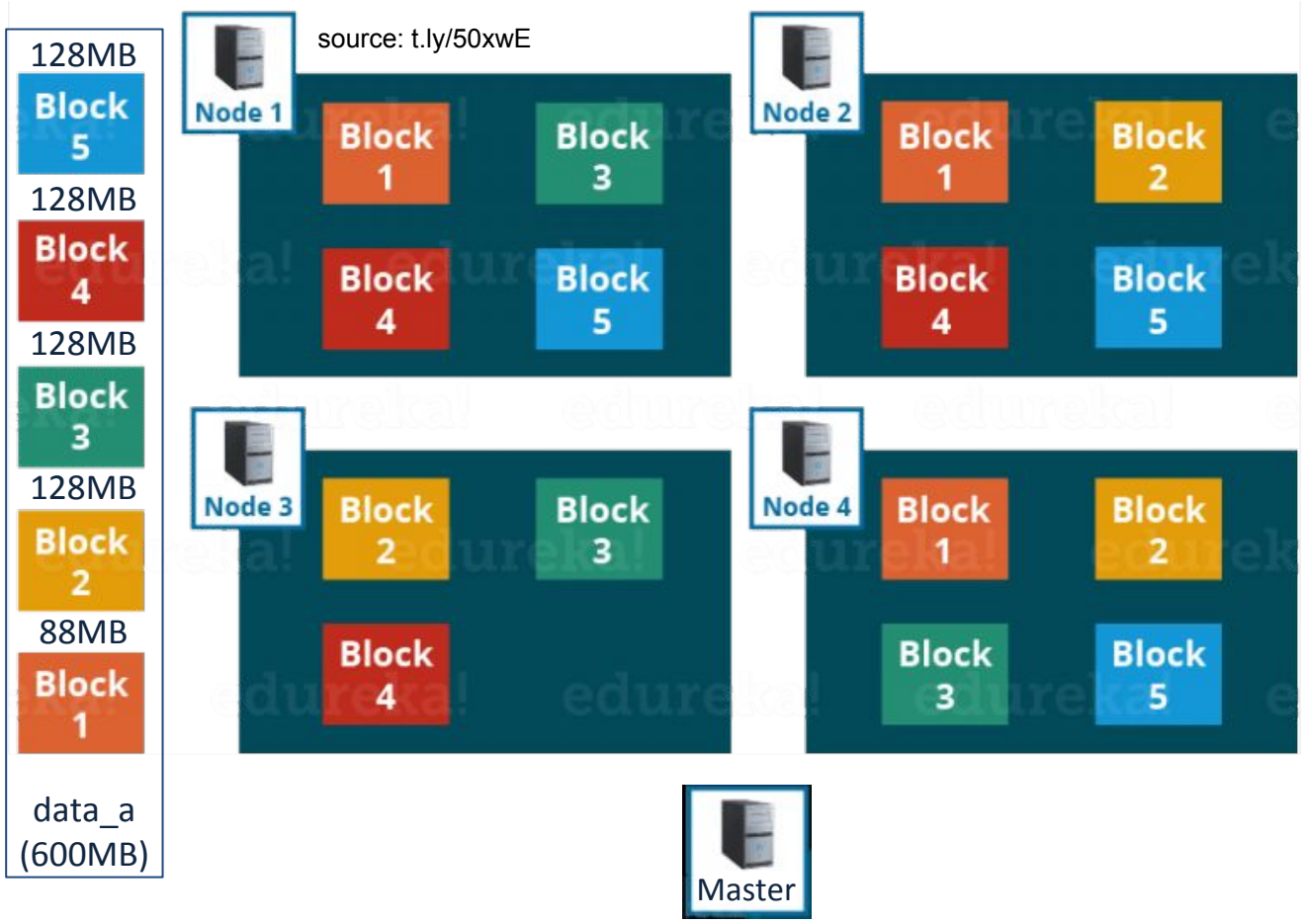
source: t.ly/50xwE

Three core components of Hadoop cluster: **storage**, processing, and resource management

- Disks are usually accessed in physical blocks, rather than a byte at a time.
- Default block size
  - Windows file system (FAT32): 512 bytes (for 8 - 32 MB) to 64 KB (for 2 - 4 GB)
  - Windows file system (NTFS): 4 KB (up to 16 TB of file)
  - Linux file system: 512 bytes
  - PostgreSQL: 8 KB
  - HDFS 1.X: 64 MB
  - **HDFS 2.X: 128 MB**
    - 15,625 times bigger than PostgreSQL
    - 250,000 times bigger than Linux file system

X 250,000 =

Three core components of Hadoop cluster:
storage, **processing**, and resource management



Example

1. **user A** requests the **data**
2. master node finds the location of blocks of the **data**
3. **user B** requests the **image_a**

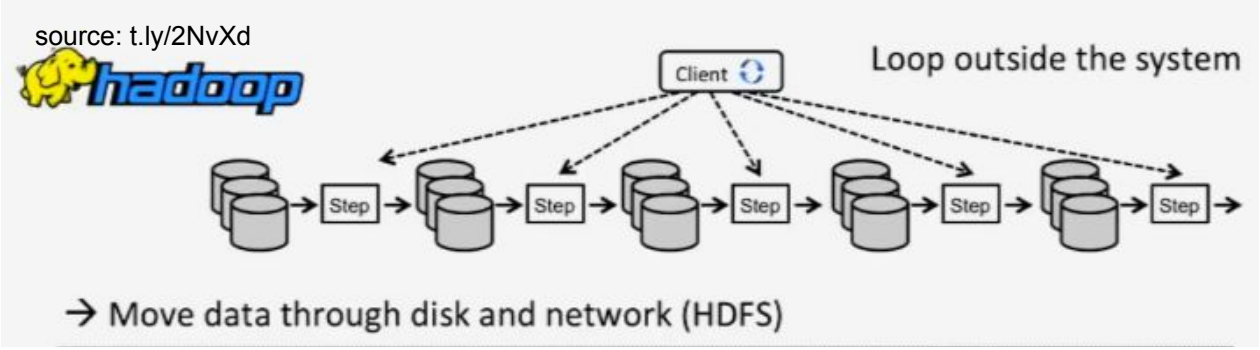
4-1. master node sends all blocks to **user A**

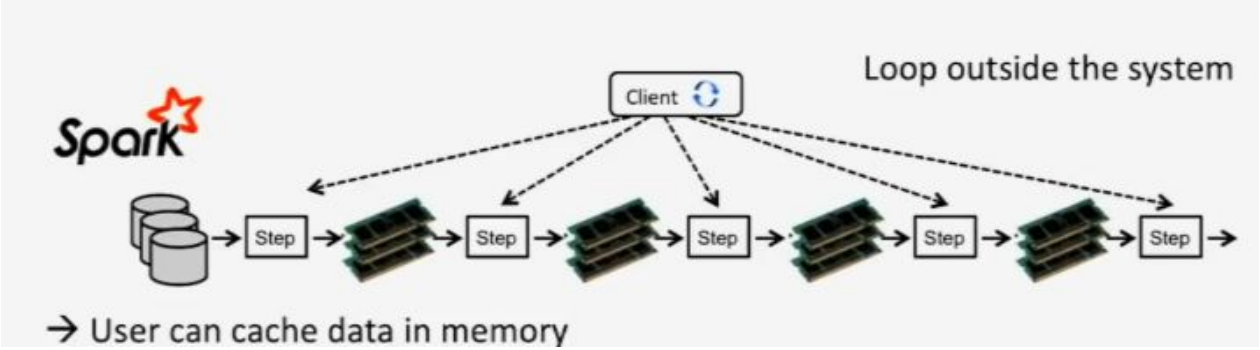
4-2. master node finds the location of blocks of the **data**

5. master node sends all data blocks to **user B**

Three core components of Hadoop cluster:
storage, **processing**, and resource management

source: t.ly/2NvXd

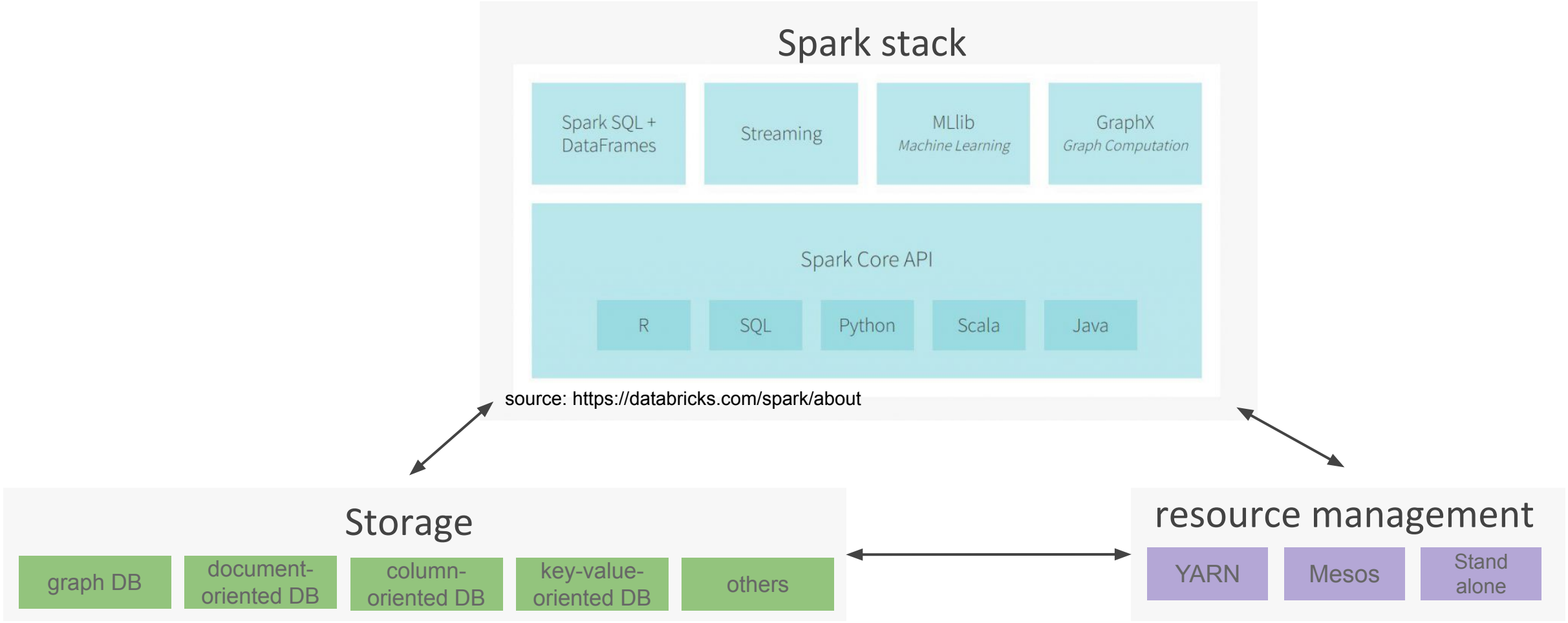




- Hadoop MapReduce is the first generation of processing engine in Hadoop ecosystem
- By now, Spark is the default choice for big data processing.
- Spark won the 2014 Daytona GraySort Benchmark. "**Spark can sort 100 TB of data 3X faster than Hadoop MapReduce on one-tenth of the machines.**"
- The **in-memory computation** is the reason for the growth of Spark's popularity

Three core components of Hadoop cluster:
storage, **processing**, and resource management

Spark stack and other components



[Demo][Hands-on] Machine Learning with PySpark

- Prescriptive analytics with clustering and SQL aggregation at scale

Takeaway

- Describe what Big data and Data science are
- Data science project needs not only data scientist
- Python is the first class citizenship in the data science world

- Data analysis with Python
- Data visualization with Python

- Describe what machine learning is
- Use cases of machine learning
- Decision tree and Clustering in depth

- Python in Hadoop ecosystem
- Distributed storage in Hadoop ecosystem
- Spark in Hadoop ecosystem

External resources
Kaggle: data science community

- The largest and most diverse data community in the world.
  (As of May 2016, Kaggle had over 536,000 registered users)
- Competitions and take an advantage of using “Kernels” (www.kaggle.com/competitions)
- Example kernels:
  - Can you improve lung cancer detection? (https://goo.gl/MV01o3)
  - Transforming How We Diagnose Heart Disease (https://goo.gl/b9Rta1)
  - Predict West Nile virus in mosquitos across the city of Chicago (https://goo.gl/VdVKtF)
  - and more

External resources
data science introduction

- Foundations of Data Science — Spring 2016 (Berkeley University)
  - https://data-8.appspot.com/sp16/
- Getting and Cleaning Data (Johns Hopkins University)
  - https://www.coursera.org/learn/data-cleaning#
- Tutorials
  - https://www.kaggle.com/kanncaa1/plotly-tutorial-for-beginners
  - https://www.kaggle.com/c/data-science-bowl-2018/kernels
- Visualization examples
  - Matplotlib (https://matplotlib.org/gallery/index.html)
  - Plotly (https://plot.ly/python/)

External resources data science

- Coursera:
  - Andrew Ng's Machine Learning
  - Natural language processing
  - Statistics One
  - Interactive programming with Python
  - Applied Machine Learning in Python
- Udacity:
  - Statistics
  - Introduction to Artificial Intelligence
- MIT
  - http://www.youtube.com/watch?v=Nx0lRBaXoz4
  - Artificial Intelligence MIT
- Khan Academy
  - https://www.khanacademy.org/math/linear - algebra

External resources
big data technology

- Docker
  - https://docs.docker.com/get-started/
- Spark
  - IBM Spark Course 1 - https://bigdatauniversity.com/courses/what - is - spark/
  - IBM Spark Course 2 - https://bigdatauniversity.com/courses/spark - rdd/
  - Spark MLlib - https://bigdatauniversity.com/courses/spark - mllib/
  - Spark GraphX - https://bigdatauniversity.com/courses/spark - graphx/
  - UC Berkeley AMP Lab camps (videos & labs) - http://ampcamp.berkeley.edu/
  - Spark Examples - http://ampcamp.berkeley.edu/big - data - mini - course/
- Kafka - IBM Kafka course - https://bigdatauniversity.com/courses/simplifyingdatapipelines/
- Flink - http://dataartisans.github.io/flink - training/index.html
- Cloudera - Online resources - http://www.cloudera.com/training/library.html ; Quick VM
- HortonWorks tutorials - http://hortonworks.com/tutorials/ ; SandBox VM

SMART on FHIR applications (https://gallery.smarthealthit.org)

- SMART® (Substitutable Medical Applications and Reusable Technologies) provides a standard for how EHR systems and their applications authenticate and integrate.
- FHIR (Fast Healthcare Interoperability Resource) is a draft standard describing data formats and resources and an application programming interface (API) for exchanging electronic health records.
- Examples:
  - Cardiac Risk (https://gallery.smarthealthit.org/app/cardiac-risk)
  - HealthDecision (https://gallery.smarthealthit.org/app/healthdecision)

Your opinion matters

- Tutorial Feedback (Just 4 questions)
- https://t.ly/MG6J1

Ready to use Python in Hadoop ecosystem?

"Humans are allergic to change. They love to say, 'We've always done it this way.' I try to fight that. That's why I have a clock on my wall that runs counter-clockwise."

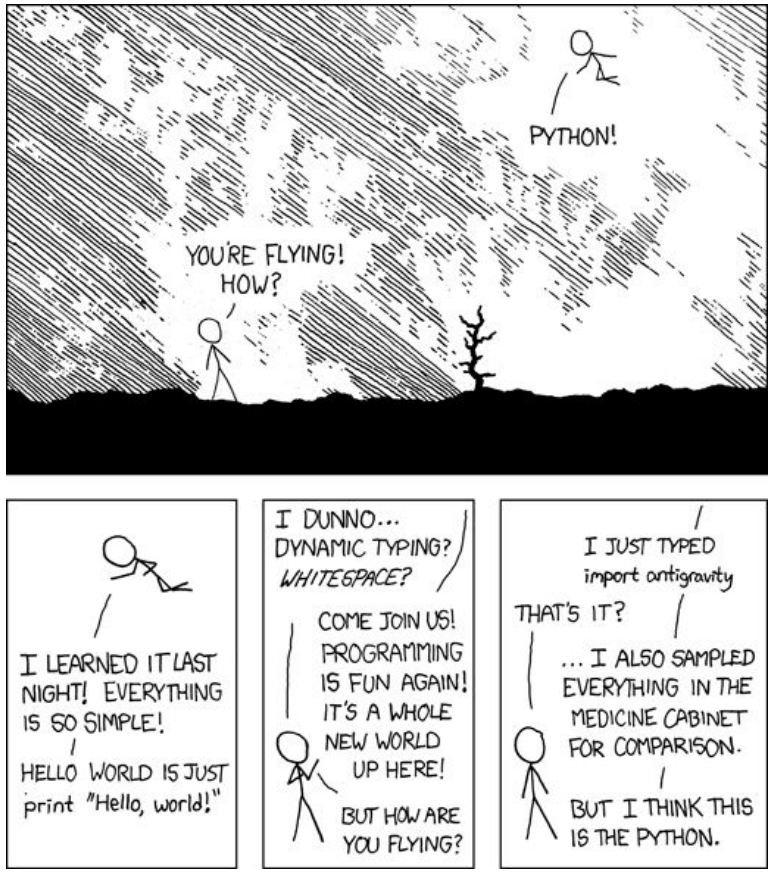
- Grace Hopper

UMM
UNIVERSITÄTSMEDIZIN
MANNHEIM

Medizinische Fakultät Mannheim
der Universität Heidelberg

Universitätsklinikum Mannheim

THANK YOU



Kim Hee and Lukas Götz
Graduate research assistant
at Heinrich-Lanz-Center (HLZ) for Digital Health