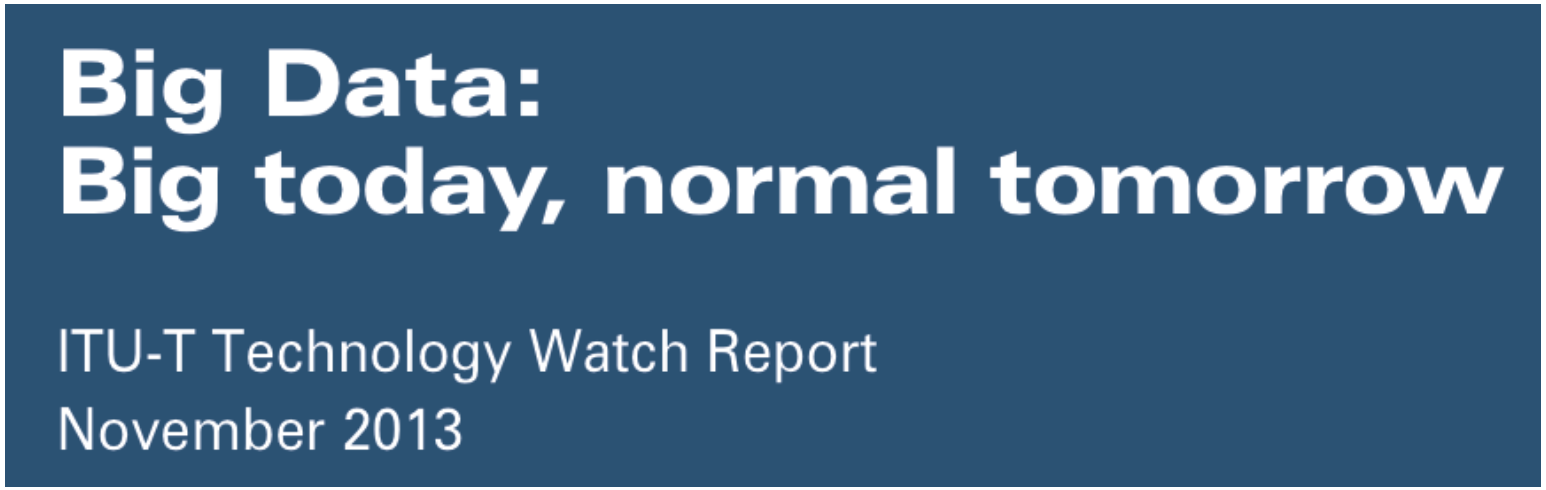
NBTC – ITU Training
on Building IoT solutions for e-applications

Session 6: IOT, Big Data and analytics

CELEBRATING 25 YEARS OF ACHIEVEMENTS

THE MEANING OF BIG?

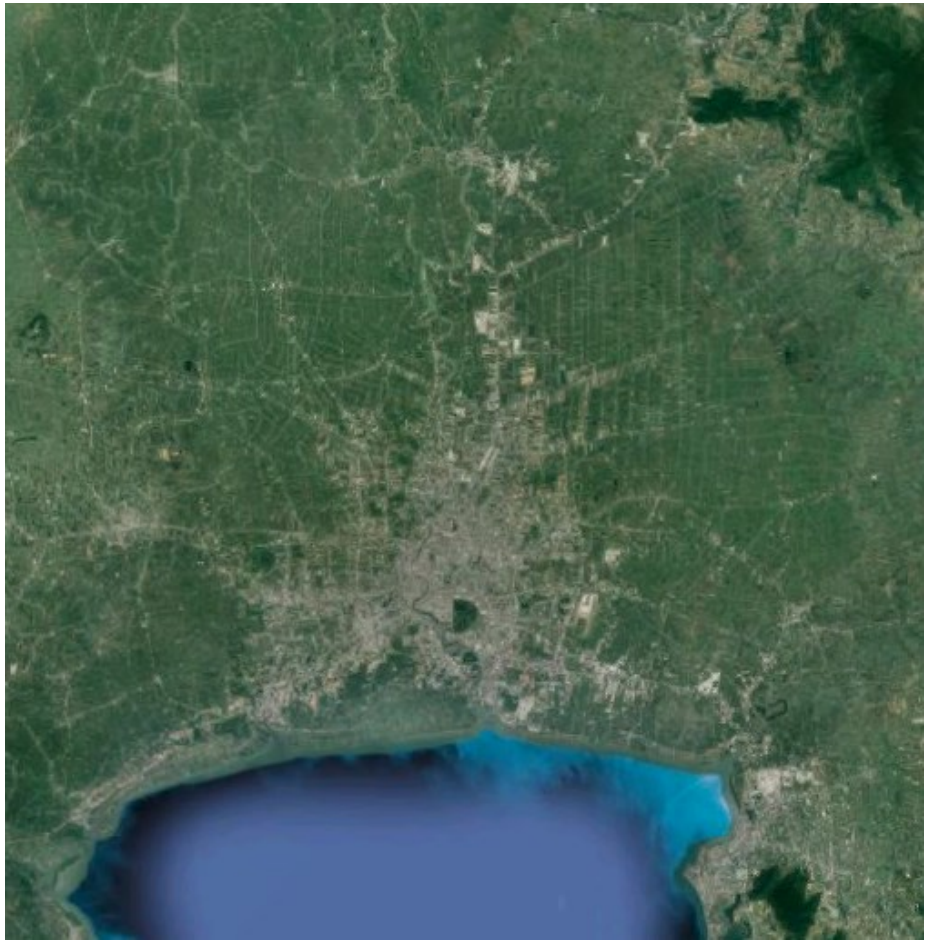


https://www.itu.int/dms_pub/itu-t/oth/23/01/T23010000220001PDFE.pdf

LET'S TRY TO MAKE IT BIG!

CELEBRATING 25 YEARS OF ACHIEVEMENTS

CASE STUDY: MONITORING AIR POLLUTION IN BANGKOK



CELEBRATING 25 YEARS OF ACHIEVEMENTS

SCENARIO 1

AREA: **1,569** **km²** ~ 40 x 40 km

SPATIAL SAMPLING: **1** **station every** **100** **meters**

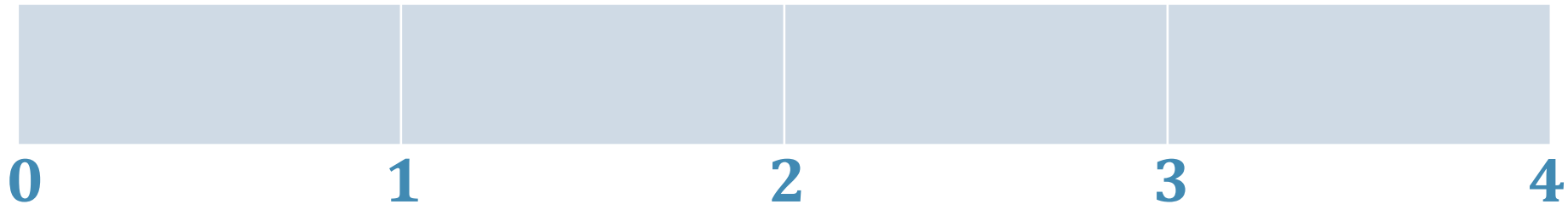
TEMPORAL SAMPLING: **1** **measurement every** **1** **hour**

DATA STRUCTURE: **~100** **bytes** **[TIME, LON, LAT, STATION_ID, CO_2, SO_2, PM, …]**

~ 376 MB / day 1,569 x (10 x 10) x 100 x 24

~ 137 GB / year 1,569 x (10 x 10) x 100 x 24 x 365

SCENARIO 1: HOW BIG IS IT?



*** NOT REALLY!**

SCENARIO 2

AREA: **1,569** **km²** ~ 40 x 40 km

SPATIAL SAMPLING: **1** **station every** **50** **meters**

TEMPORAL SAMPLING: **1** **measurement every** **1** **minute**

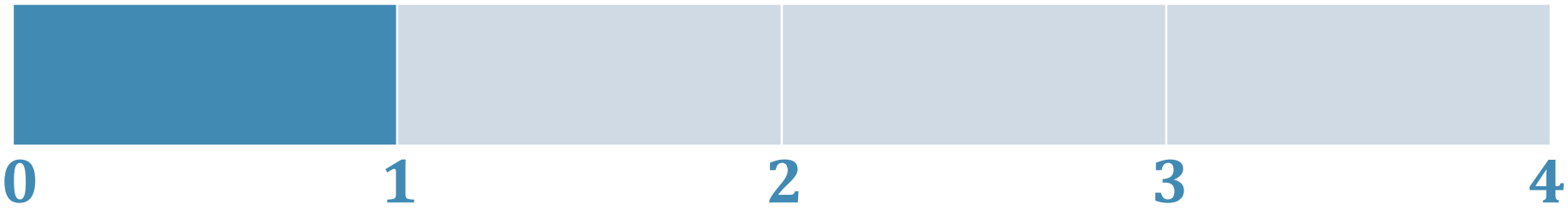
DATA STRUCTURE: **~1000** **bytes/measurement**

[TIME, LON, LAT, STATION_ID, CO_2, SO_2, PM, … , … , … , … , …]

~ 903 GB / day 1,569 x (20 x 20) x 24 x 60 X 1000

~ 329 TB / year 1,569 x (20 x 20) x 24 x 60 x 1000 x 365

SCENARIO 2: HOW BIG IS IT?



Volum

- *Only use case justify accessing a year of data*

SCENARIO 3

AREA: **1,569** **km²** ~ 40 x 40 km

SPATIAL SAMPLING: **1** **station every** **50** **meters**

TEMPORAL SAMPLING: **1** **measurement every** **1** **second**

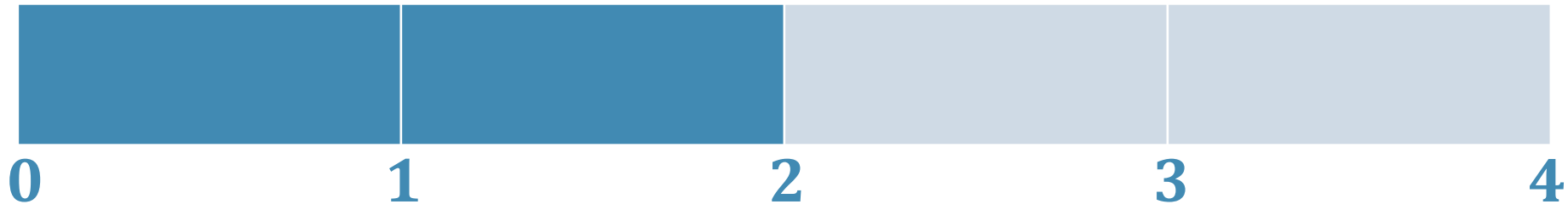
DATA STRUCTURE: **~1000** **bytes/measurement**

[TIME, LON, LAT, STATION_ID, CO_2, SO_2, PM, … , … , … , … , …]

~ 54 TB / day 1,569 x (20 x 20) x 24 x 3600 X 1000

~ 20 PB / year 1,569 x (20 x 20) x 24 x 3600 x 1000 x 365

SCENARIO 3: HOW BIG IS IT?



Volume

Velocity

** of acquisition and processing*

SCENARIO 4

IDEM AS SCENARIO 3 ~54 TB / day and ~20 PB /year

+ CROWD-SOURCED DATA - citizen science, third party institutions, …

+ WEB APP. DATA COLLECTION - perception on air quality (good, moderate, poor)

+ SENTIMENT ANALYSIS ON SOCIAL NETWORKS

+ IMAGE CLASSIFICATION (SATELLITE IMAGERY, CAMERAS, …)

SCENARIO 4: HOW BIG IS IT?

Volum
e

Velocity

Variety

Veracity

…

CHECK LIST

Different problems | different solutions

CELEBRATING 25 YEARS OF ACHIEVEMENTS

USE CASE FIRST *

- **WHAT IS THE USE CASE ?**
- **WHAT DECISION WE WANT TO MAKE ?**
- **WHICH DATA WILL SUPPORT THAT DECISION ?**

* *as opposed to* ***"let's collect everything we can, then we will see what we can do with it"*** *syndrom.*

DOMAIN KNOWLEDGE IS A KEY INPUT

- DOMAIN KNOWLEDGE PROVIDES PERSPECTIVE AND INSIGHTS
- MIGHT DOWNSIZE CONSIDERABLY THE AMOUNT OF DATA NEEDED

REAL-TIME vs. BATCH PROCESSING

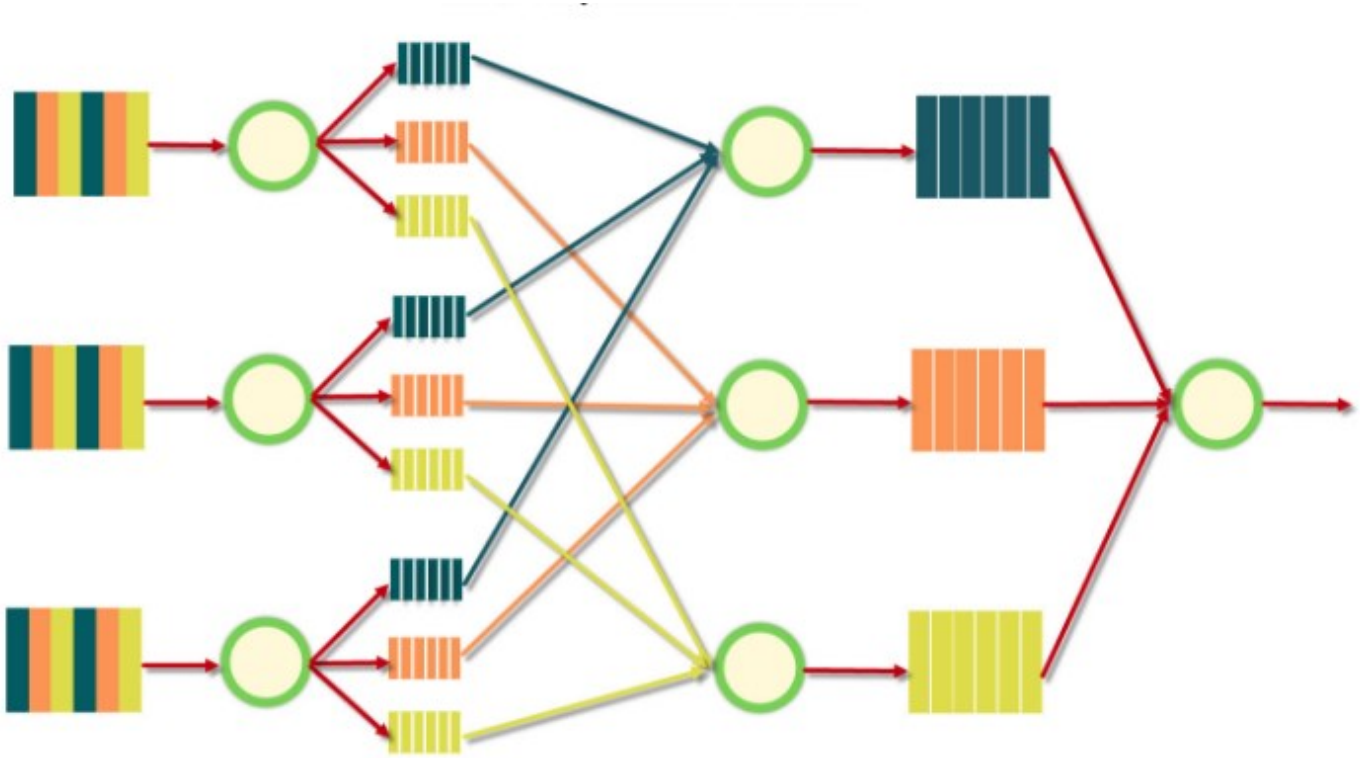
- DOES YOUR USE CASE REQUIRE PROCESSING HISTORICAL DATA REAL-TIME ?
- DOES CLASSICAL DRILL-DOWN ROLL-UP STRATEGY ADDRESS YOUR PROBLEM ?

TECHNOLOGY OVERVIEW

Evading the hype

CELEBRATING 25 YEARS OF ACHIEVEMENTS

CLASSICAL DIVIDE & CONQUER APPROACH



https://blog.sqlauthority.com/2013/10/09/big-data-buzz-words-what-is-mapreduce-day-7-of-21/

USE CASE: # OF PARTICIPANTS BY MOBILE OS USED

KEY: ANDROID

VALUE: 12

KEY: iOS

VALUE: 6

...

KEY: OTHERS

VALUE: 3

NAIVE IMPLEMENTATION

VS.

DIVIDE & CONQUER | PARALLELIZING | MapReduce*

*ROLE PLAYING: MAKING CONRETE THE MAP-SHUFFLE-REDUCE PHASES

KEY: ANDROID

VALUE:

KEY: iOS

VALUE:

KEY: OTHERS

VALUE:

*TO BE PRINTED IN 10 COPIES

CELEBRATING 25 YEARS OF ACHIEVEMENTS

WHAT ABOUT VELOCITY, VARIETY, … ?

- BATCH vs. STREAM PROCESSING
- VARIETY OF DBMS TECHNOLOGIES
- PIPELINES (DATA MOVING AROUND)
- VARIETY OF PROGRAMMING PARADIGMS
- SCALABILITY

CANONICAL TECHNOLOGICAL ECOSYSTEM/STACK

- **CLUSTERED FILE SYSTEM:** HDFS, GFS, …
- **“DIVIDE & CONQUER”:** Hadoop, Spark, …
- **“FLAT FILE STORAGE” | API:** Simple Storate Service (S3), …
- **RDBMS:** PostGres, …
- **NoSQL DB:** MongoDB, DynamoDB, …
- **“PREPARE DATA FOR DATA ANALYTICS” | DATA WAREHOUSE:** Redshift, …
- **“MOVING DATA AROUND”:** AWS Data pipeline, …
- **STREAMING PROCESSING:** AWS Kinesis, Spark stream, …
- **BI/ANALYTICS CLIENT PLATFORM**: JasperSoft, Python, R, SAS, Tableau…

LEARNING RESOURCES

- **COURSERA:** https://www.coursera.org/courses?languages=en&query=big+data
- **PLURALSIGHT:** https://www.pluralsight.com/search?q=big%20data
- **UDACITY:** https://www.udacity.com/courses/all
- **…**

CELEBRATING 25 YEARS OF ACHIEVEMENTS

THANK YOU

CELEBRATING 25 YEARS OF ACHIEVEMENTS