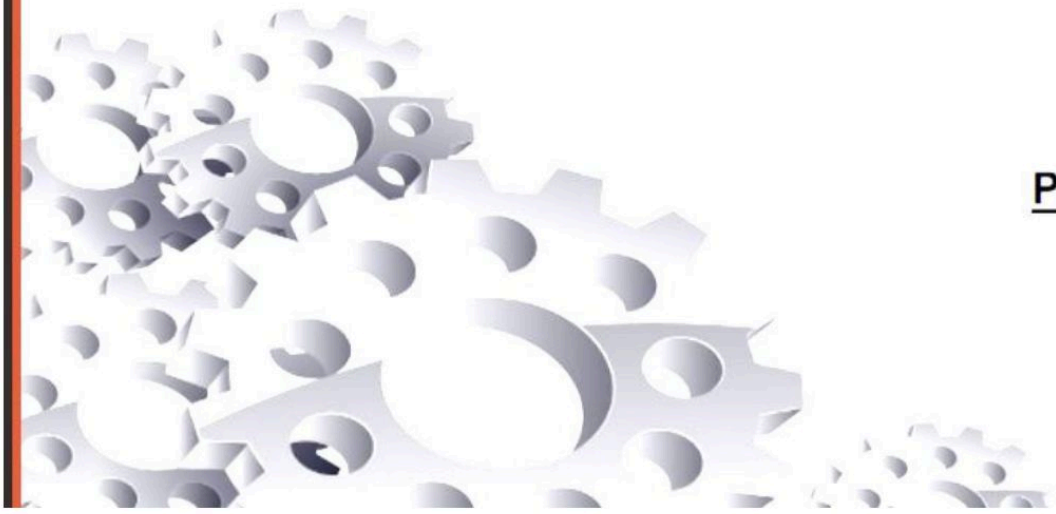
BIG DATA ANALYSIS WITH IBM CLOUD DATABASES



PRESENTED BY:

P.SHARMILA

M.VANITHA

A.VIMALADEVI

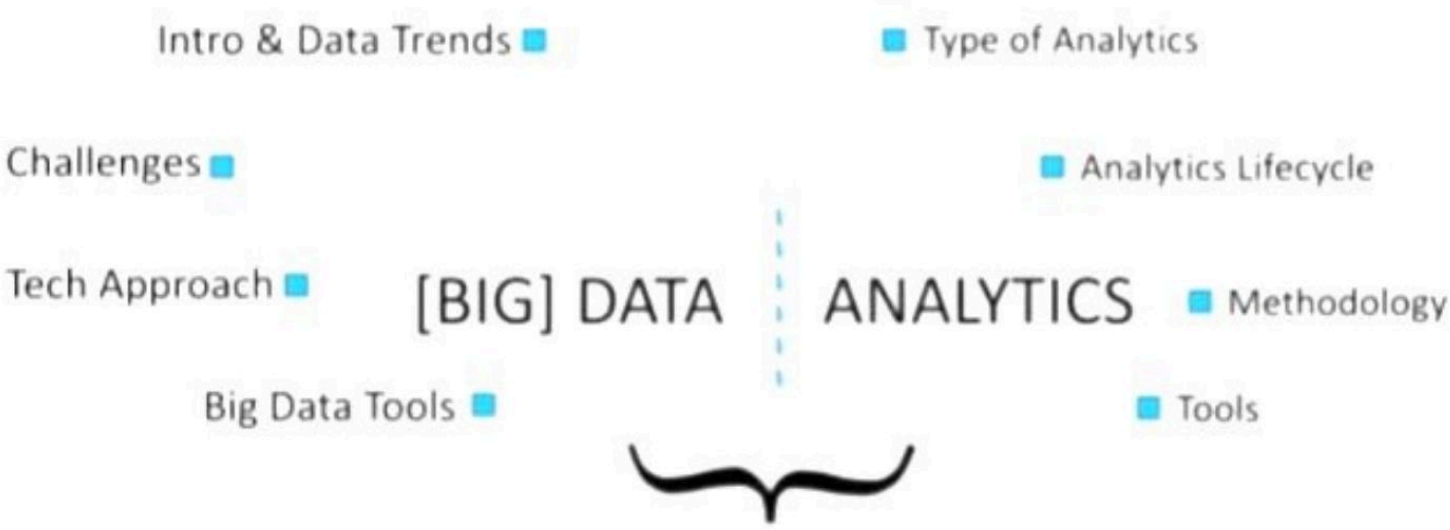
ABOUT ME

Currently work in Telkomsel as senior data analyst

8 years professional experience with 4 years in big dat and predictive analytics field in telecommunication industry

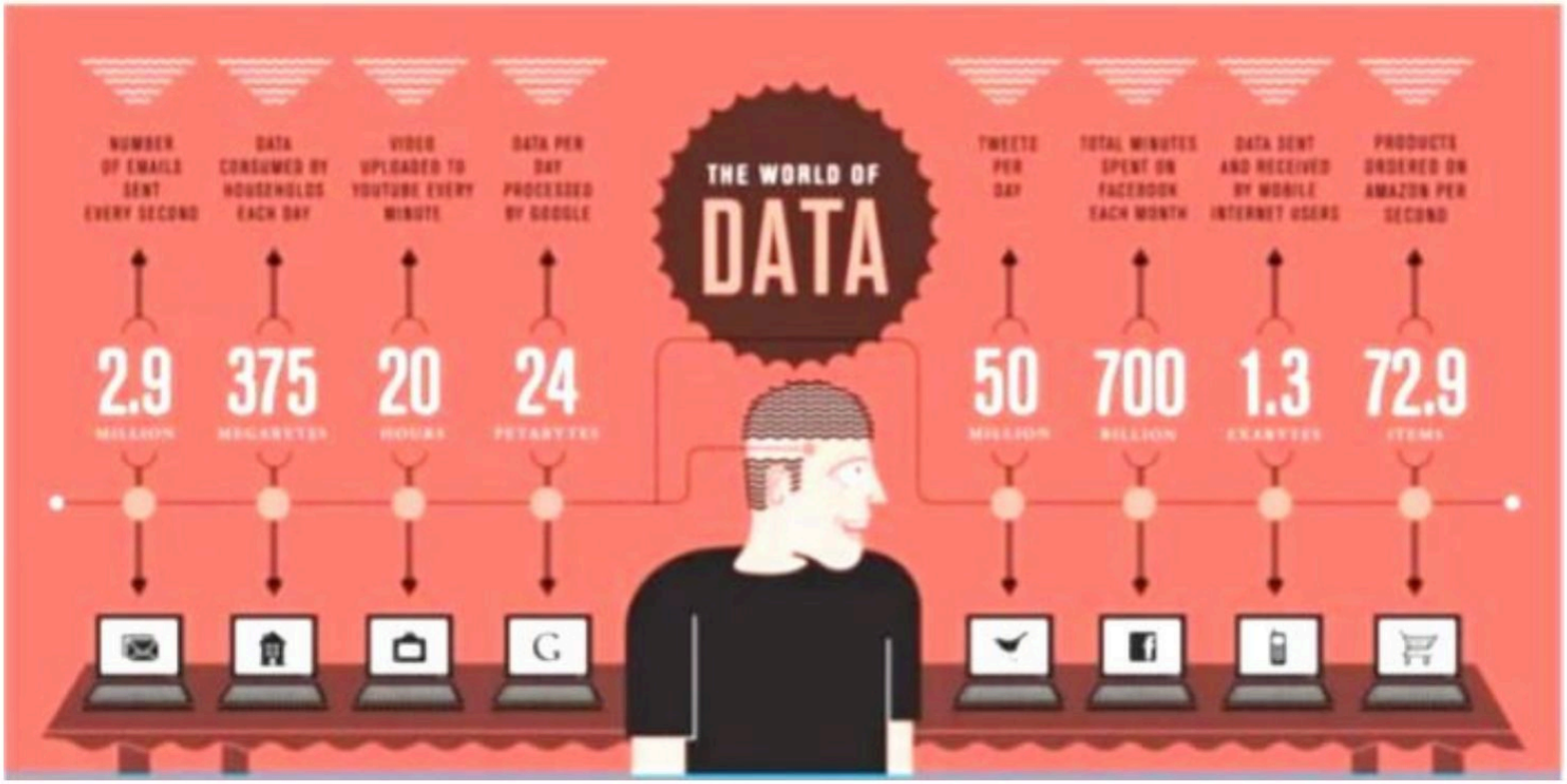
Bachelor from Computer Science, Gadjah Mada University & get master degree from Magister of Information Technology, Universitas Indonesia

WHAT'S IN THIS SLIDE



- Use Cases (Sentiment Analysis)
  - What's Trending
  - Where to Start

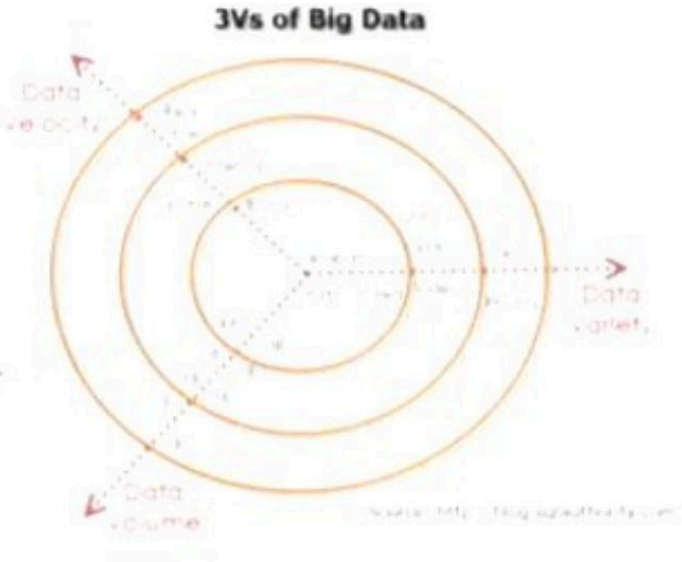
THE WORLD OF DATA



DATA VS BIG DATA

Big data is just data with:

- More **volume**
- Faster data generation (**velocity**)
- Multiple data format (**variety**)



World's data volume to grow 40% per year & 50 times by 2020 [1]

Data coming from various human & machine activity

CHALLENGES

- More data = more storage space
  - More storage = more money to spend (RDBMS server needs very costly storage)
- Data coming faster
  - Speed up data processing or we'll have backlog
- Needs to handle various data structure
  - How do we put JSON data format in standard RDBMS?
  - Hey, we also have XML format from other sources
  - Other system give us compressed data in gzip format
- Agile business requirement.
  - On initial discussion, they only need 10 information, now they ask for 25? Can we do that? We only put that 10 in our database
  - Our standard ETL process can't handle this

STORAGE COST

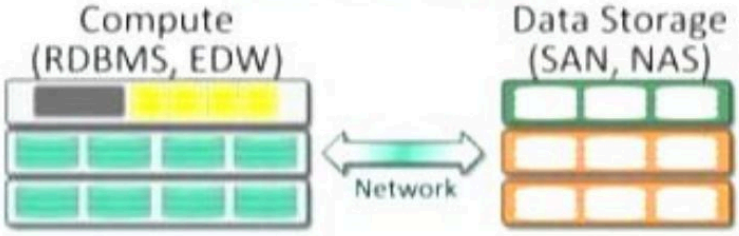
In Terms of storage cost, Hadoop has lower comparing to standard RDBMS.

Hadoop provides highly scalable storage and process with fraction of the EDW Cost

	Hadoop	Netezza	Exadata	Extreme Data Appliance (1650)
Cost / Terabyte	**$333**	**$10,000**	**$14,000**	**$16,500***
Hadoop Benefit		30x saving	42x saving	50x saving

STORAGE & COMPUTE TOGETHER

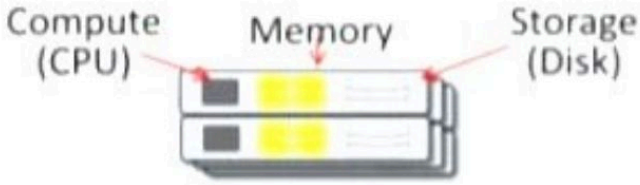
The Old Way



Expensive, Special purpose, "Reliable" Servers
Expensive Licensed Software

- Hard to scale
- Network is a bottleneck
- Only handles relational data
- Difficult to add new fields & data types

The Hadoop Way

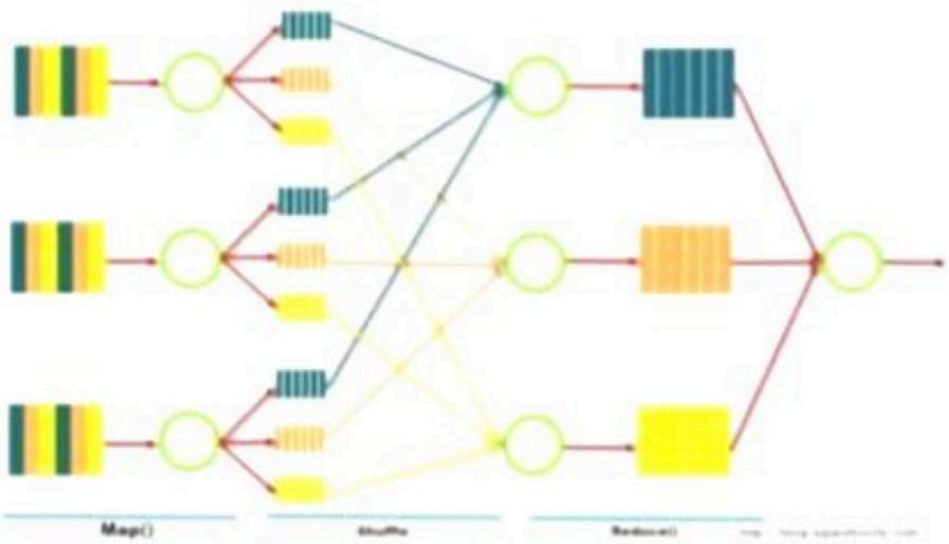


Commodity "Unreliable" Servers
Hybrid Open Source Software

- Scales out forever
- No bottlenecks
- Easy to ingest any data
- Agile data access

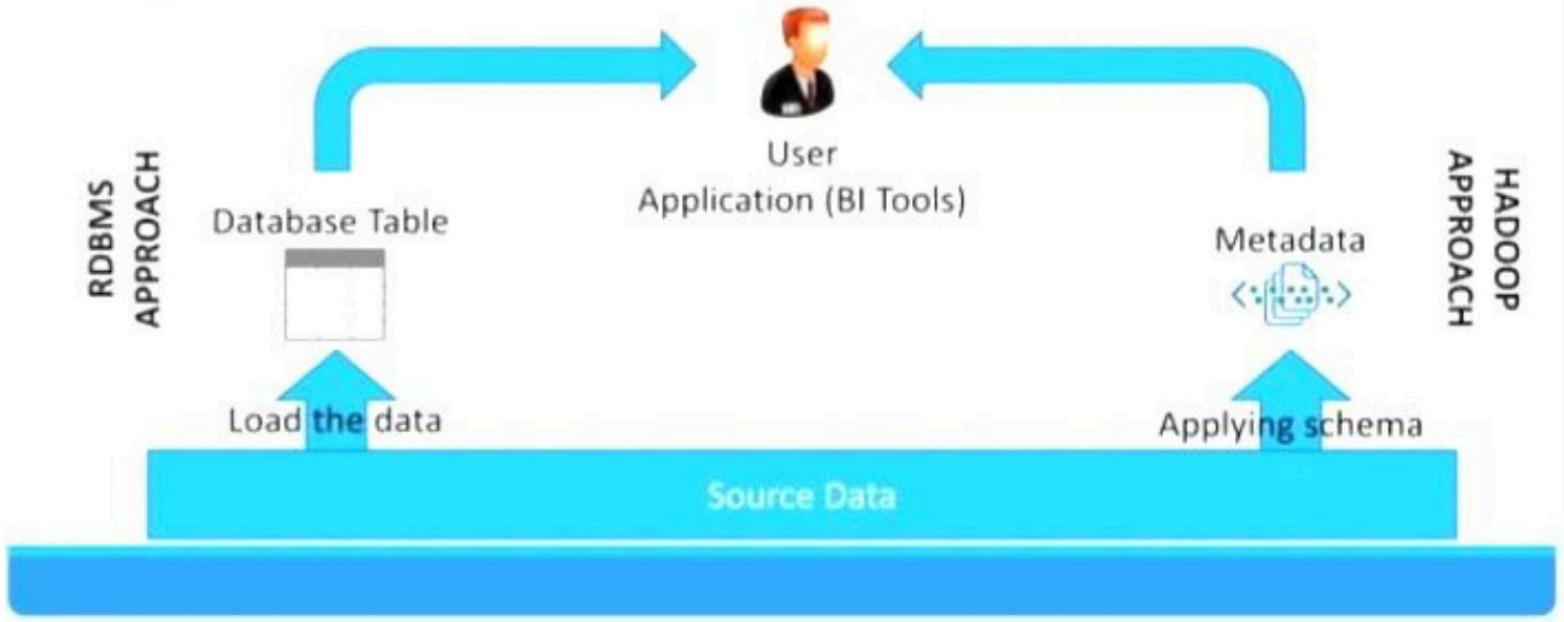
MAP REDUCE APPROACH

- Process data in parallel way using distributed algorithm on a cluster
- Map procedure performs filtering and sorting data locally
- Reduce procedure performs a summary operation (count, sum, average, etc.)



HADOOP vs UNSTRUCTURED DATA

- Hadoop has HDFS (Hadoop Distributed File System)
- It is just file system, so what you need is just drop the file there ☺
- Schema on read concept



HIVE

- The Apache Hive ™ data warehouse software facilitates querying and managing large datasets residing in distributed storage.
- With Hive you can write the schema for the data in HDFS
- Hive provide many library that enable you to read various data type like XML, JSON, or even compressed format
- You can create your own data parser with Java language
- Hive support SQL language to read from your data
- Hive will convert your SQL into Java MapReduce code, and run it in cluster

- Apache spark is fast and general engine for large-scale data processing
- Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk
- You can write spark application in Java, Scala, Python, or R
- Spark support library to run SQL, streaming, and complex analysis like graph computation and machine learning
- https://spark.apache.org/

ANALYTICS

ANALYTICS IS IN YOUR BLOOD

- Do you realize that you do analytics everyday?
- I need to go to campus faster!
- Hmm.. Looking at the sky today, I think it'll be rain
- Based on my mid term and assignment score, I need to get at least 80 in my final exam to pass this course
- I stalked her social media. I think she is single because most of her post only about food :p

DESCRIPTIVE & PREDICTIVE

- Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that, for example, patterns might emerge from the data.
  - In Information System Design course, most of the student get C grade (11 people). There is 4 people get A, 7 get B, 7 get D, and 7 get E
  - Fulan only post his activity on Facebook at weekend
- Predictive analytics is the branch of data mining concerned with the prediction of future probabilities and trends.
- The central element of predictive analytics is the predictor, a variable that can be measured for an individual or other entity to predict future behavior.
  - Fulan should be has a job. Because he always left home at 7 in the morning and get back at 6 afternoon

PREDICTIVE ANALYTICS

There is 2 types of predictive analytics:

Supervised

Supervised analytics is when we know the truth about something in the past

Example:

we have historical weather data. The temperature, humidity, cloud density and weather type (rain, cloudy, or sunny). Then we can predict today weather based on temp, humidity, and cloud density today

Machine learning to be used: Regression, decision tree, SVM, ANN, etc.

Unsupervised

Unsupervised is when we don't know the truth about something in the past. The result is segment that we need to interpret

Example:

We want to do segmentation over the student based on the historical exam score, attendance, and late history

APPLYING THE CONTEXT

Source *Raw & unstructured*

IMEI & TAC
Device info


Location


URL access


CDR


Socmed data, Complaint, Survey

+

Context


Point Of Interest, sentiment library, socmed buzzer, website category

Derived Information


Commute pattern
Hangout location


Idols
Political view
Pain point


Community leader
Family member
Communication spending

ANALYTICS LIFECYCLE

Business Understanding

- Gathering problem information
- Defining the goal to solve the problem
- Defining expected output
- Defining hypothesis
- Defining analysis methodology
- Measuring the business value

Data Understanding

- Define variables to support hypothesis
- Cleaning & transforming the data
- Create longitudinal data/trend data
- Ingesting additional data if needed
- Build analytical data mart

Data Mining & Modeling

- Defining target variable
- Splitting data for training and validating the model
- Defining analysis time frame for training and validation
- Correlation analysis and variable selection
- Selecting right data mining algorithm
- Do validation by measuring accuracy, sensitivity, and model lift
- Data mining and modeling is an iterative process

ANALYTICS LIFECYCLE

Model Interpretation

- Describe the importance of each variable
- Visualize overall model by creating decision tree for example
- Define business action based on the model result

Model Operationalization

- Define the model scoring period
- Integrate model result with execution system (campaign system, CRM, etc)
- Create operational process that timely, consistent, and efficient

Model Monitoring

- Create monitoring process for model evaluation
- Evaluate the model based on real-world result
- Monitor and evaluate the business impact

Analytics and modeling is an iterative process. Data model will become obsolete and need to evolve to accommodate changes in behavior

BUILDING THE METHODOLOGY

Analysis Domain

- What is the analysis domain? Is it for male only? Is it for housewife or worker? Your "customer" segment has different behavior

Type of Analysis

- Do we need only descriptive analysis? Or we need to go with predictive analysis?

Supervised or Unsupervised?

- Do we need to build unsupervised clustering/segmentation for this analysis?

Define Analysis Time Window

- What time window of data we need for behavior observation?
- What is the prediction time window?
- Is there any seasonal event on that time window?

ANALYTICS TOOLS

Microsoft Excel. Very powerful tools to do statistical data manipulation, pivoting, even doing simple prediction

SQL is just the language. Your data lying in database? SQL will help to filter, aggregate and extract your data

RapidMiner provide built-in RDBMS connector, parser for common data format (csv, xml), data manipulation, and many machine learning algorithm. We can also create our own library. Latest version of RapidMiner can connect to Hadoop and do more complex analysis like text mining. Free version is available (community edition)

KNIME. Known as a powerful tools to do predictive analytics. Overall function is similar to RapidMiner. Latest version of KNIME can connect to Hadoop and do more complex analysis such as text mining. Free version is available

Tableau is one of the famous tools to build visualization on top of the data. Tableau also powerful to create interactive dashboard. Free version is available with some limitation

QlikView. Similar to Tableau, QlikView designed to enable data analyst to develop a dashboard or just simple visualization on top of the data. Free version is available