BIG DATA Analytics

Quick Start Training

July 2018

Labs247
Technology in motion

CHAPTER 01

Big Data Overview

The Rise of Big Data

- Technology Growth
- Internet Adoption
- People Behaviour
- Digitize Everything
- **Competition**

3S Problem

What is Big Data ?

Buzz phrase, no single definition

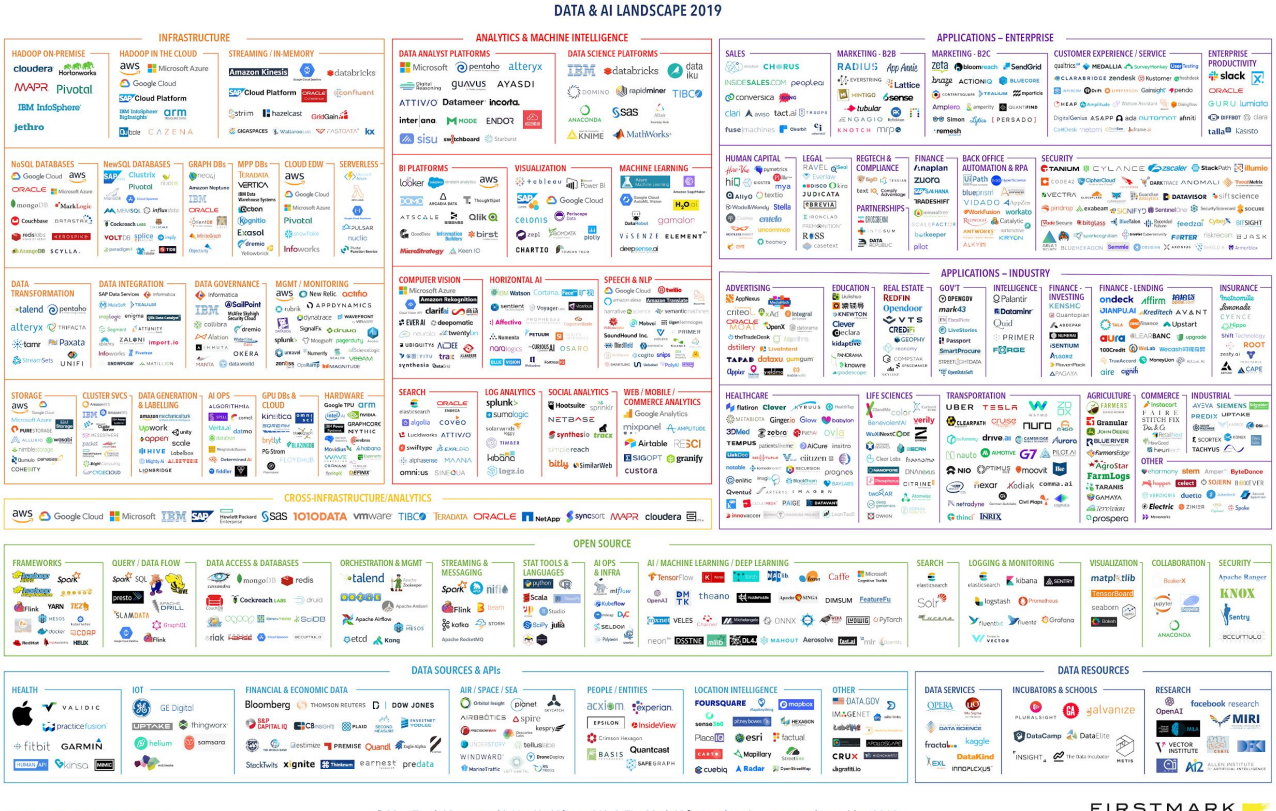
Wikipedia

> Big data is a term for data sets that are **so large or complex** that traditional data processing application software is **inadequate to deal with them**. Challenges include **capture, storage, analysis, data curation, search, sharing, transfer, visualization, querying, updating and information privacy**.

Now, refer to Big Data Analytics

Matt Turck

VC at FirstMark Capital. Previously, Managing Director at Bloomberg Ventures and co-founder of TripleHop Technologies. Occasional angel investor. Startup mentor (Techstars, DreamIt, ERA, FGVN, NYC Venture Fellows). Organizer of two large monthly tech community events, Data Driven NYC and Hardwired NYC



source:
https://mattturck.com/data2019

Big Data and AI Landscape

Matt Turck
VC at FirstMark Capital
mattturck.com/bigdata2019

INFRASTRUCTURE

Infrastructure

Hadoop on Prem, Hadoop on Cloud, Streaming/In Memory, NoSQL Databases, NewSQL Databases, Graph DBs, MPP DBs, Cloud EDW, Data Transformation, Data Integration, Data Governance, Mgmt/Monitoring, Storage, Cluster Svcs, App Dev, Crowd Sourcing, Hardware, GPU DBs

ANALYTICS

Analytics

Data Analytics Platforms, Data Science Platforms, BI Platforms, Visualization, Machine Learning, Computer Vision, Horizontal AI, Speech & NLP, Search, Log Analytics, Social Analytics, Web/Mobile/Commerce Analytics

APPLICATIONS – ENTERPRISE

Application Enterprise

Sales - Marketing B2B - Marketing B2C - Customer Service - Human Capital - Legal - Finance - Enterprise Productivity - Back Office Automation - Security

APPLICATIONS – INDUSTRY

Application Industry

Advertising - Education - Government
Finance - Real Estate - Insurance - Healthcare
Life Science - Transportation - Agriculture
Commerce - etc

CROSS-INFRASTRUCTURE/ANALYTICS

Cross Infrastructure/Analytics

OPEN SOURCE

Open Source

Frameworks, Query/Data Flow, Data Access, Coordination, Streaming, Stat Tools, AI/ML/DL, Search, Logging & Monitoring, Visualization, Collaboration, Security

DATA SOURCES & APIS

Data Source & APIS

Health, IoT, Financial & Economic Data, Air/Space/Sea, People, Location Intelligence, Other

DATA RESOURCES

Data Source

Open Source Technology

- Most of big data component is open source
- We can download the code, use and modify freely
- Require adequate human resources
- Lots of choices

Disruptive Technology

Since it first appearance to the current, big data has changed the existing and established business model.

- Open source – zero license
- Proven by big internet company
- Active community
- Fast adoption

Proven by Big Internet Company

- Invented by internet company
- Used in production environment
- Shared to open source community
- Specific function

Active Community

- Many developers are involved in technology development
- Supported and sponsored by big company

DATA is the new OIL

We don't have better algorithms,
we just have more data

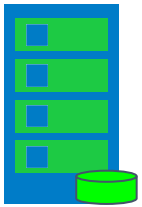
- Peter Norvig - Director of Research at Google

CHAPTER 02

YAVA Data Management Platform

Parallel Processing

→ **Read 1 TB Data**

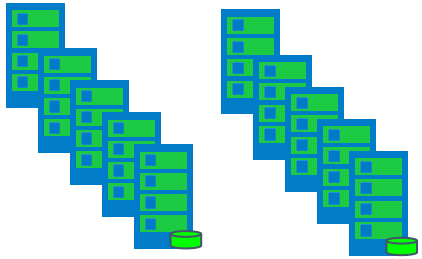


1 Machine

- 4 I/O channels
- each channel : 100 MB/s

Theoretically : **45 minutes**

- 250 GB per channel



10 Machine

- 4 I/O channels
- each channel : 100 MB/s

Theoretically : **4.5 minutes**

- 25 GB per channel

Parallel Processing is Complex

- Job balancing
- Hardware failure and failover
- Most analysis tasks need to be able to combine the data

What is Hadoop ?

- Open Source Platform for data management
- Combination of distributed storage and distributed processing
- Computer cluster built from commodity hardware
- Framework written in java programming
- Offering scalability and high performance
- The name Hadoop is not an acronym; Doug Cutting named it after his son's toy elephant

History of Hadoop

- Mike Cafarella and Doug Cutting started the Nutch project in 2002
- In 2003, Google published Google File System paper, that described the architecture of Google's distributed file system
- By adopting GFS, Nutch Distributed File System (NDFS) began to be implemented on the Nutch project in 2004
- In 2004, Google published the paper that introduced MapReduce to the world
- Early in 2005, the Nutch developers had a working MapReduce implementation in Nutch
- In February 2006 they moved out of Nutch to form an independent subproject of Lucene called Hadoop
- April 2006 – Hadoop 0.1.0 was released

Who use Hadoop ?

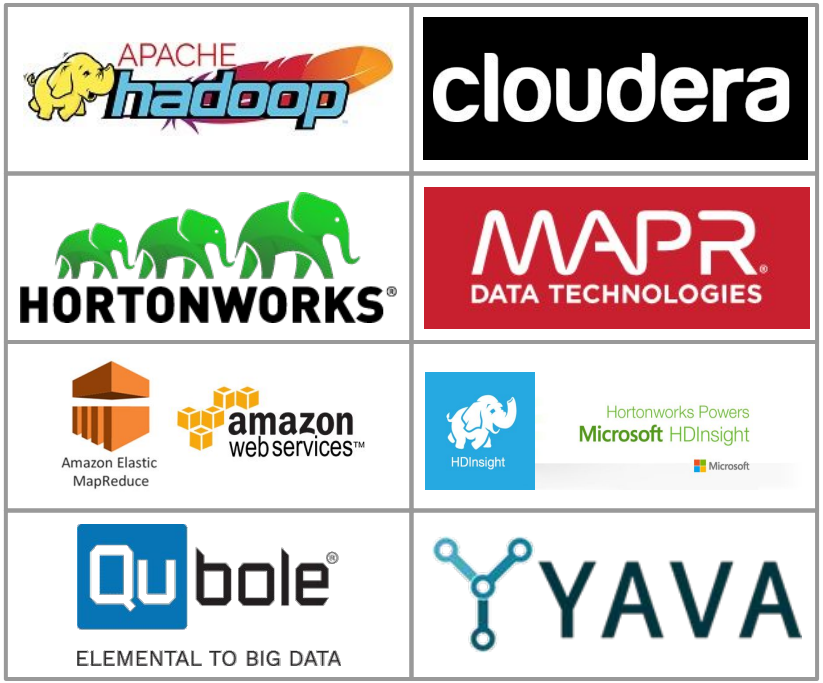
In 2008, Yahoo! Inc. the world's largest Hadoop production application, runs on a Linux cluster with more than 10,000 cores. Now more than 100,000 CPUs in > 40,000 computers running Hadoop.

In 2010, Facebook claimed that they had the largest Hadoop cluster in the world with 21 PB of storage, In June 2012, they announced the data had grown to 100 PB, now the data grow 0.5 PB every day

Spotify, 1650 node cluster : 43,000 virtualized cores, ~70TB RAM, ~65 PB storage

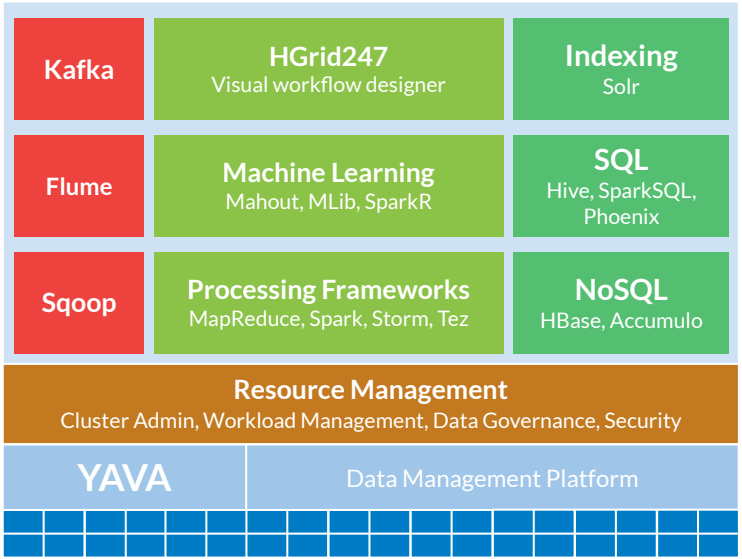
Source : wiki.apache.org/hadoop/PoweredBy

Hadoop Distribution



Magic Quadrant for Data Management Solutions for Analytics

YAVA Data Management Platform



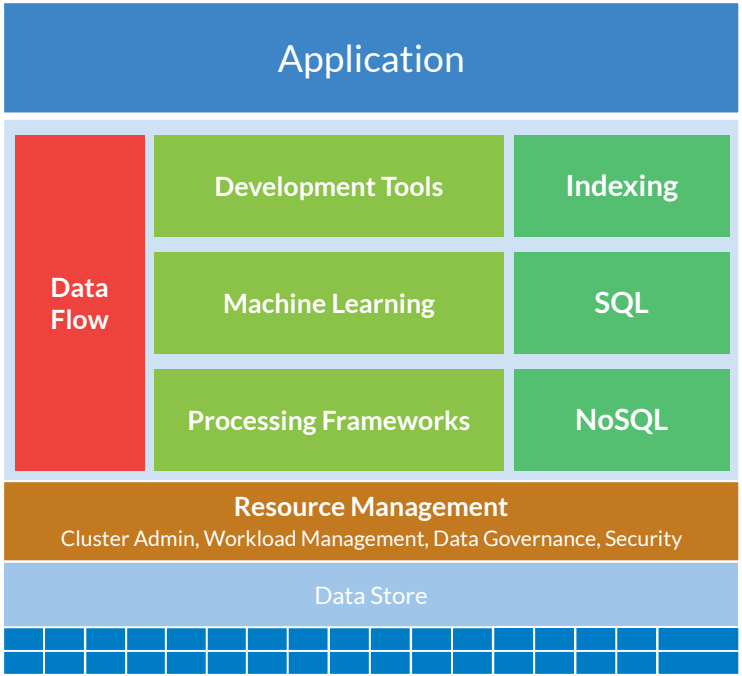
All in one data management platform
Programming/Scripting :

- Java, Python, Scala, R
- SQL
- HGrid247 - Visual Designer

For further info : **yava.labs247.id**

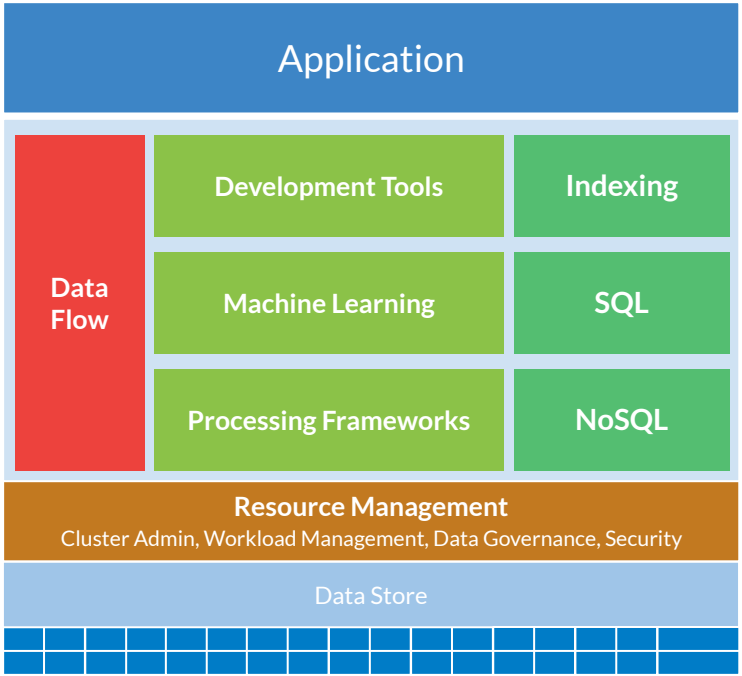
Big data and artificial intelligence platform based on open source component. It is designed to make organization easier to implement big data.

Yava Component



Data Store	HDFS, GlusterFS
Cluster Management	Ambari
Data Governance	Atlas, Falcon
Security	Knox, Ranger
Workload Management	Yarn, Zookeeper, OOzie, Slider
Data Flow	Flume, Sqoop, Kafka, NiFi
Processing Frameworks	MapReduce, Spark, Storm, Tez
Machine Learning	Mahout, MlLib, SparkR, H2O
Development Tools	HGrid247, Zeppelin, Jupyter

Yava Component



NoSQL	HBase, Accumulo
SQL	Hive, Phoenix, Hawq
Indexing	Solr, Elasticsearch
Visualization	Kibana, Superset
Application	Chantel, Medan, Metron

Usage

Archival and Storage

- Retain years of data
- Retain intermediate format

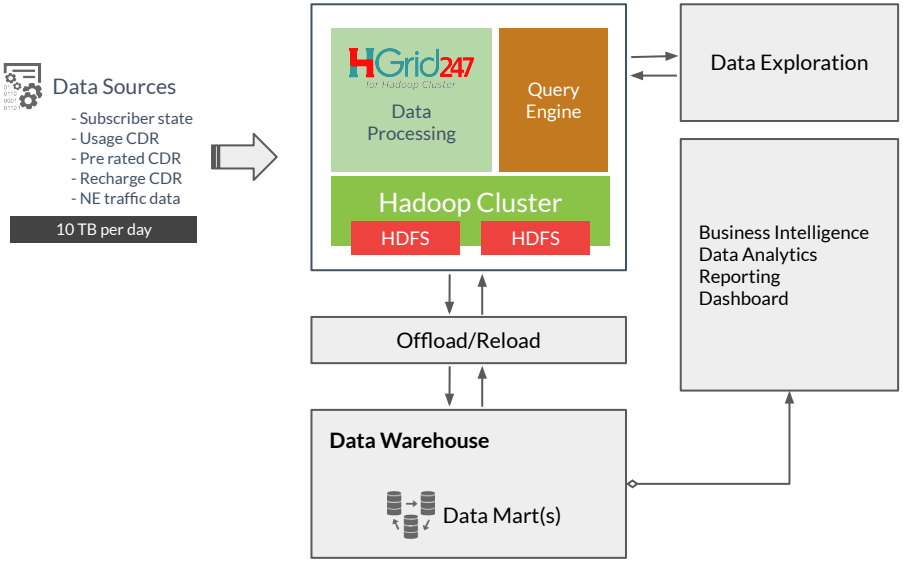
Transformation

- Map inputs and outputs where needed
- Turn unstructured data into structured at runtime

Analysis

- Explore data in-place
- Execute arbitrary code

Use Case : Data Warehouse



- Data Warehouse as a single point of truth
- Problem :
  - Cannot achieve SLA
  - Hi Cost

DATA LAKE

Data Engineering

- The data professionals who prepare the “big data” infrastructure to be analyzed by Data Scientists
- Software engineers who design, build, integrate data from various resources, and manage big data
- Run some ETL (Extract, Transform and Load) on top of big datasets and create big data warehouses that can be used for reporting or analysis by data scientists
- Typically not expected to know any machine learning or analytics for big data.
- Skills and tools: Hadoop, MapReduce, Hive, Pig, MySQL, MongoDB, Cassandra, Data streaming, NoSQL, SQL, programming.

Source : bigdatauniversity.com/blog/data-scientist-vs-data-engineer

BI Developers

- Data experts that interact more closely with internal stakeholders to understand the reporting needs, and then to collect requirements, design, and build BI and reporting solutions for the company
- Work with databases, both relational and multidimensional, and should have great SQL development skills to integrate data from different resources
- Skills and tools: ETL, developing reports, OLAP, cubes, web intelligence, business objects design, Tableau, dashboard tools, SQL, SSAS, SSIS

Source : bigdatauniversity.com/blog/data-scientist-vs-data-engineer

Data Analyst

- Experienced data professionals in their organization who can query and process data, provide reports, summarize and visualize data
- Strong understanding of how to leverage existing tools and methods to solve a problem, and help people from across the company understand specific queries with ad-hoc reports and charts
- Expected to have the mathematical or research background to develop new algorithms for specific problems.
- Skills and Tools: Data Analysts need to have a baseline understanding of some core skills: statistics, data munging, data visualization, exploratory data analysis, Microsoft Excel, SPSS, SPSS Modeler, SAS, SAS Miner, SQL, Microsoft Access, Tableau, SSAS.

Source : bigdatauniversity.com/blog/data-scientist-vs-data-engineer

Data Scientist

- The alchemist of the 21st century: someone who can turn raw data into purified insights.
- Apply statistics, machine learning and analytic approaches to solve critical business problems.
- Expected to have strong programming skills, an ability to design new algorithms, handle big data, with some expertise in the domain knowledge.
- Also expected to interpret and eloquently deliver the results of their findings, by visualization techniques, building data science apps, or narrating interesting stories about the solutions to their data (business) problems.
- Skills and tools: Python, R, Scala, Apache Spark, Hadoop, data mining tools and algorithms, machine learning, statistics

Source : bigdatauniversity.com/blog/data-scientist-vs-data-engineer

idBigData - Komunitas Big Data Indonesia

idBigData.com

IDBigData

idBigData

@idBigData

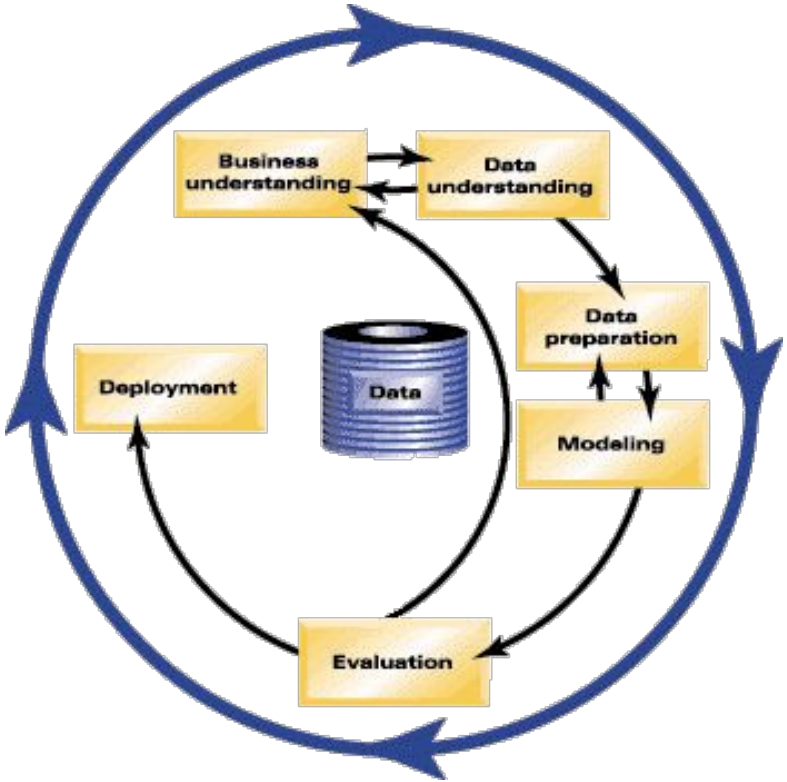
hub.idBigData.com

s.id/idbigdata

CHAPTER 03

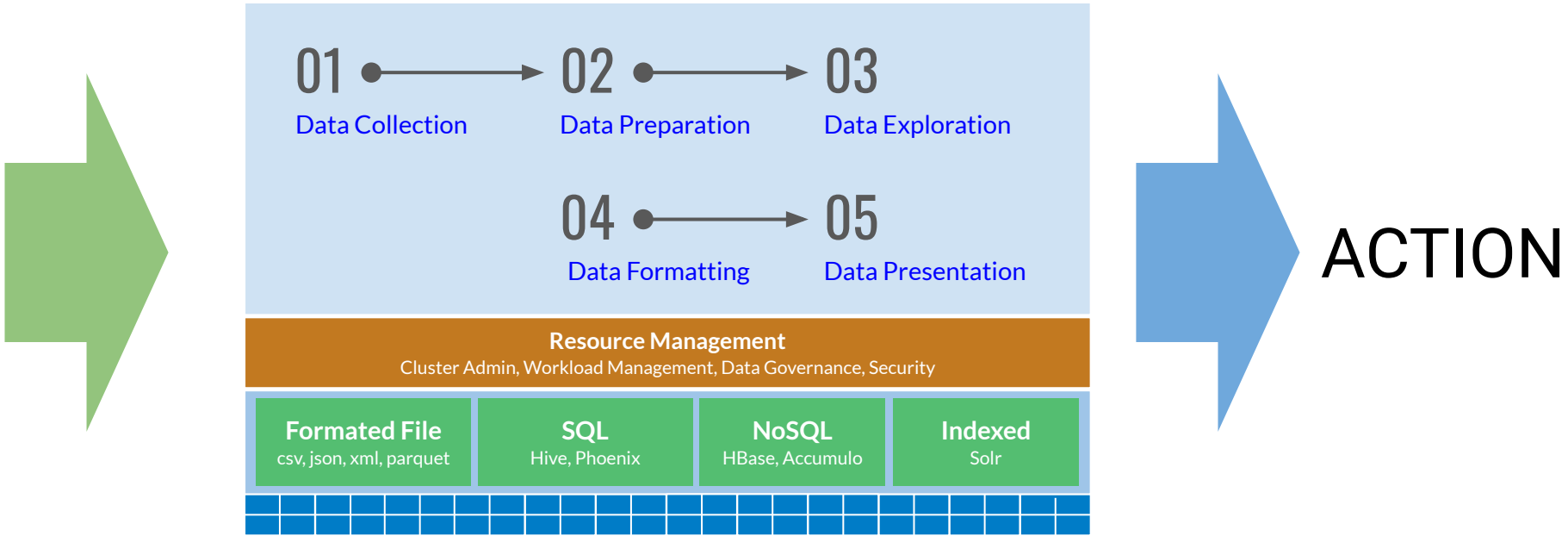
Big Data Analytics Workflow

CRISP DM Framework



- Cross-Industry Standard Process for Data Mining, is an industry-proven way to guide your data mining efforts
- Business Understanding : where the business problem is defined and characterised
- Data Understanding and Data Preparation **consume 85% of the total project time**

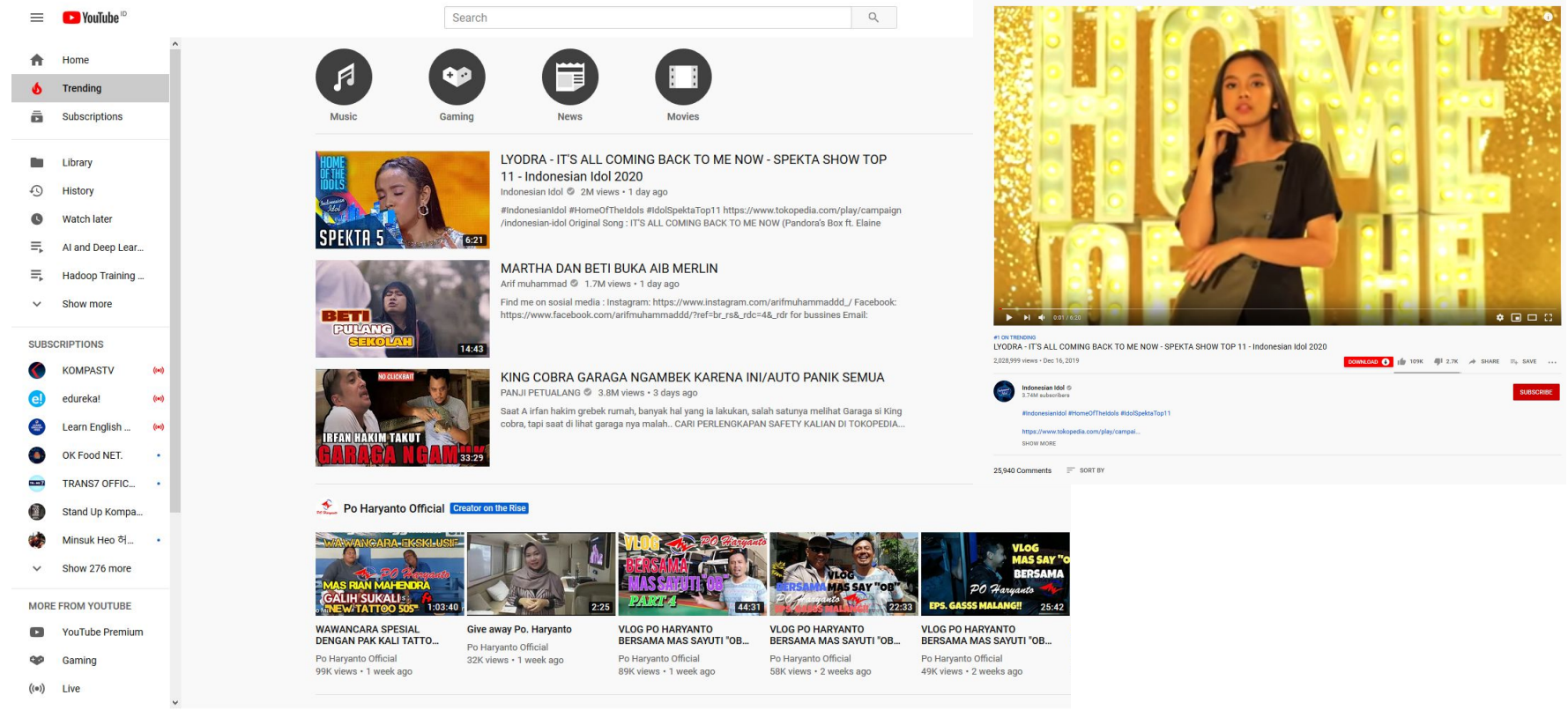
Data Journey



CHAPTER 04

Getting Closer With Business

YouTube Video Trending



Business Question

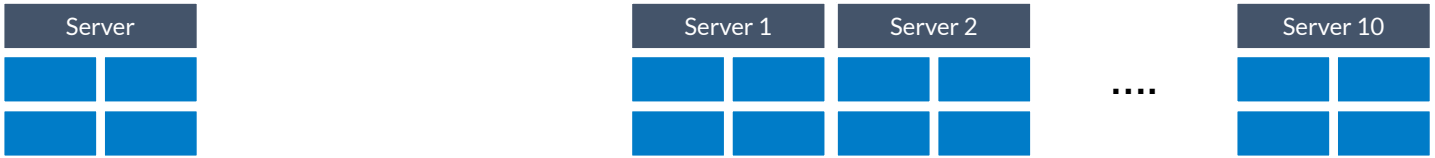
- How long usually a video can trend ?
- How many likes, dislikes, views and comments get ?
- Correlation of trending video in between countries
- Videos from which category has longer trend?
- Users like videos from which CATEGORY the most?
- What is the ratio of Likes-Dislikes and Views-Comments in different categories?

CHAPTER 05

Ingesting Data Into HDFS

Distributed File System

READ 1 TB of Data



1 Machine

- 4 I/O channel
- 100 MB/s per channel

Theoretically

- 250 GB per channel
- 45 minutes

10 Machine

- 4 I/O channel
- 100 MB/s per channel

Theoretically

- 25 GB per channel
- 4.5 minutes

Hadoop Component

Main Hadoop Component :

1. **Hadoop Distributed File System**
   a distributed file system designed to run on commodity hardware

2. **MapReduce**
   a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

3. **Yarn**
   the resource management layer for the Apache Hadoop ecosystem that allows multiple data processing engines such as interactive SQL, real-time streaming, data science and batch processing to handle data stored in a single platform

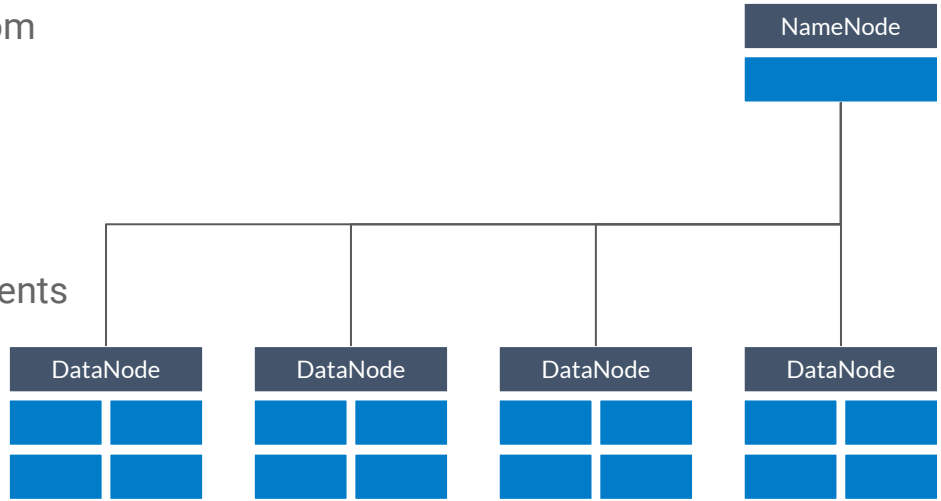
Master Slave Architecture

NameNode

1. Master service
2. Maintain and manage DataNodes
3. Records metadata i.e file size, location of blocks stored, permission, hierarchy, etc
4. Receives heartbeat and block report from DataNodes

DataNode

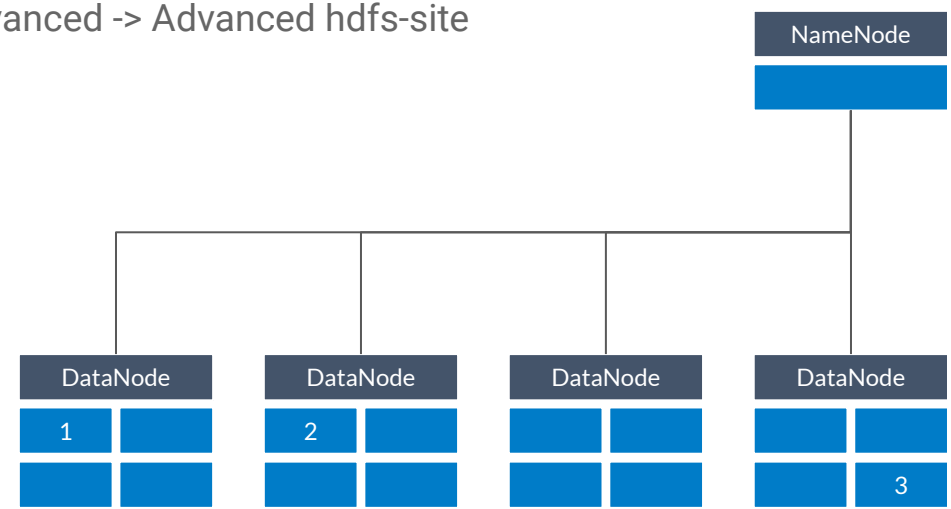
1. Slave service
2. Stores physical data
3. Serves read and write requests from clients



HDFS Block

- HDFS splits huge files into small chunks known as data blocks
- We (client and admin) do not have any control over the data block like block location
- Default size of each block is 128 MB (64 MB in Hadoop 1.x)
- Configuration : HDFS -> Configs -> Advanced -> Advanced hdfs-site

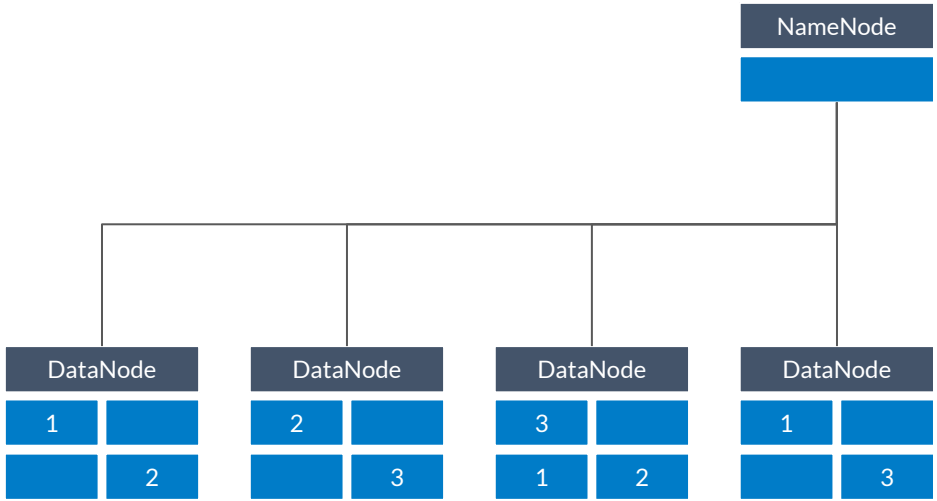
  dfs.blocksize = 134217728 (in byte)
- Sample.txt → file size : 320 MB
  - Block 1 : 128 MB
  - Block 2 : 128 MB
  - Block 3 : 64 MB



HDFS Replication

- Each block will be replicated
- Default replication is 3
- Configuration : HDFS -> Configs -> Advanced -> General

  dfs.replication = 3
- Sample.txt → file size : 320 MB
  - Block 1 : 128 MB
  - Block 2 : 128 MB
  - Block 3 : 64 MB



LABS 01: Basic Linux Command

1. Open notebook LABS01: Basic Linux Command
2. Follow the instructure

CODING TIME

LABS 02: Upload Data Into HDFS

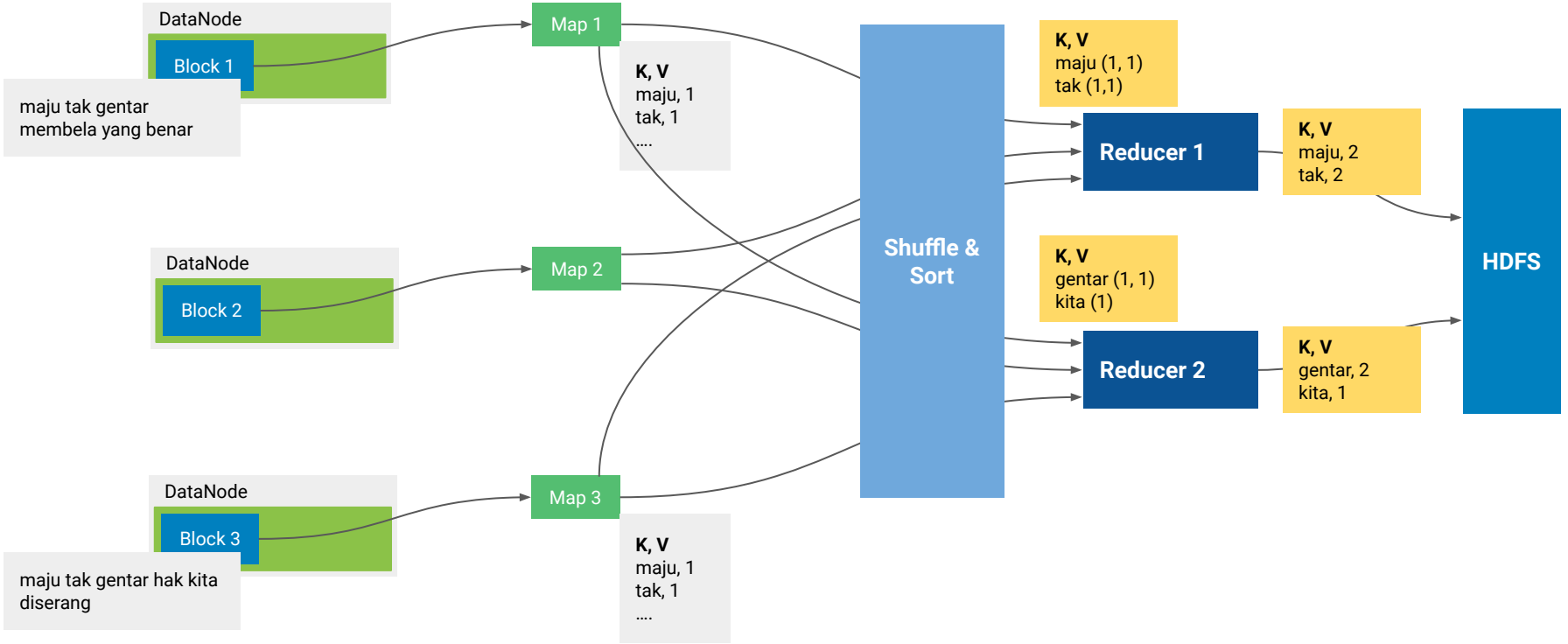
1. Open notebook LABS 01: Upload Data Into HDFS
2. Follow the instructure

CODING TIME

Yet Another Resource Negotiator (YARN)

- Introduced in Hadoop 2.x.
- Allows different data processing engines like graph processing, interactive processing, stream processing as well as batch processing to run and process data stored in HDFS
- Manage Resources:
  - scheduling
  - resources assignments (CPU and memory)
- Component:
  - Resource Manager
  - Node Manager (one per worker node)
  - Application Master (one per application)

MapReduce



LABS 03: WordCount

1. Open notebook LABS 03: WordCount
2. Follow the instructure

CODING TIME

Terasort Benchmark

- The purpose of Terasort is to test the CPU/Memory power of the cluster and to sort 1TB of data by the a 10-byte ASCII key in the shortest amount of time possible.
- This benchmark provides combined testing of the HDFS and MapReduce layers of a Hadoop cluster.
- The application is included in hadoop distribution

LABS 04: Terasort Benchmark

1. Open notebook LABS 04: Terasort Benchmark
2. Follow the instructure

CODING TIME

CHAPTER 06

Exploring Data With Hive

What is Apache Hive

hive.apache.org : **data warehouse software** that facilitates reading, writing, and managing large datasets residing in distributed storage using SQL

Wikipedia : a **data warehouse infrastructure built on top of Hadoop** for providing data summarization, query, and analysis

It provides :

- Tools to enable easy access to data via SQL, thus enabling data warehousing tasks such as extract/transform/load (ETL), reporting, and data analysis
- A mechanism to impose structure on a variety of data formats
- Access to files stored either directly in Apache HDFS™ or in other data storage systems such as Apache HBase
- Query execution via Apache Tez™, Apache Spark™, or MapReduce

Hive is intended for data analysts who familiar with SQL

Hive is NOT ….

a relational database

- Hive uses a database to store metadata, but the data that Hive processes is stored in HDFS

designed for OLTP

- Hive runs on Hadoop
- Therefore, latency for Hive queries is generally high (even for small jobs)

suited for real-time queries and row-level updates

- Hive is best used for batch jobs over large sets of immutable data (such as web logs)

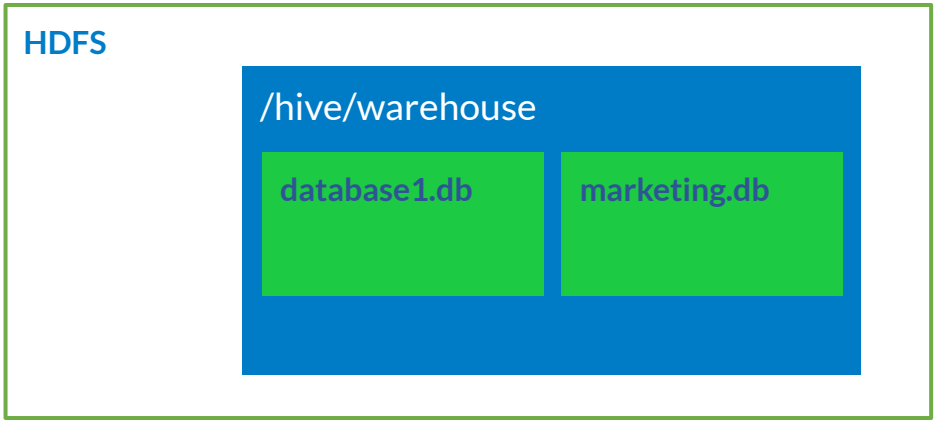
Schema On Read

At the time of loading into the hive, the data is not validated.
While in conventional database, the data is validated in accordance with a scheme that has been defined, if the data does not fit the scheme, it will be rejected.

Scheme is applied at the time of reading data
You can load the data before you know what to do with it, it offers you the ability to store structured, unlawful, and/or data that are not organized

Database

- A simply abstraction to group tables and other data unit
- To avoid naming conflicts for tables, views, partitions, columns, and so on



Create Database Example

```
hive> CREATE DATABASE mydatabase;


hive> CREATE DATABASE IF NOT EXISTS test_db
          COMMENT "Test Database created for tutorial"
          WITH DBPROPERTIES(
               'Date' = '2016-08-22',
               'Creator' = 'Bob',
               'Email' = 'my@email.com'
          );
```

Show Database

SHOW DATABASE|SCHEMA [LIKE identifier_with_wildcards**]**;

- **LIKE** – Optional. Allows us to filter the database names using a regular expression.

EXAMPLE :

```
hive> show databases;
OK
default
mydatabase
test_db
Time taken: 0.072 seconds, Fetched: 3 row(s)
hive> SHOW DATABASES LIKE '*db*';
OK
test_db
Time taken: 0.014 seconds, Fetched: 1 row(s)
```

Misc. Database Command

- Describe Database
  (DESCRIBE|DESC) DATABASE|SCHEMA [EXTENDED] database_name;

- Alter Database
  ALTER DATABASE|SCHEMA database_name **SET DBPROPERTIES** (property_name=property_value, ...);

- **ALTER (DATABASE|SCHEMA)** database_name **SET OWNER [USER|ROLE]** user_or_role;

- Use Database
  USE database_name;

- Drop Database
  DROP (DATABASE|SCHEMA) [IF EXISTS] database_name **[RESTRICT|CASCADE];**

Table

- A collection of related columns
- Can be filtered, projected, joined and unioned
- There are 2 types of tables

➔ **Managed tables**
managed by Hive by moving data into its warehouse directory
if tables are dropped, both data and metadata (schema) are deleted

➔ **External tables**
tables data will not be copied into hive warehouse directory
if tables are dropped only the schema from metastore will be deleted but not the data files from external location

Create Table

Example command to create a table :

- CREATE TABLE mytable01
  (c1 string, c2 float, c3 list<map<string, struct<p1:int, p2:int>>);

- HDFS directory of the table:
  /apps/hive/warehouse/mytable01

External Table

- Data stored in existing file in HDFS
- Tables and partition can be created
- File format must be in Hive-compatible format
- On dropping table, only the metadata drops

- Example command used to create an external table

```
CREATE EXTERNAL TABLE t_external
(c1 string, c2 int)
LOCATION ‘user/file/mydata’;
```

LABS 05: Create Database

1. Open notebook LABS 05: Create Database
2. Follow the instructure

CODING TIME

Youtube Dataset

1	video_id	Video id - 11 char
2	trending_date	Date
3	title	Video title
4	channel_title	Channel name
5	category_id	Category_id related to category reference
6	publish_time	Date time video first published
7	tags	Tag, multiple tag, separated by '|'
8	views	Number of view - int
9	likes	Number of likes - int

10	dislikes	Number of dislikes - int
11	comment_count	Number of user comment - int
12	thumbnail_link	Url of thumbnail
13	comments_disabled	
14	ratings_disabled	
15	video_error_or_removed	
16	description	Video description

Format: text file separated by comma

LABS 06: Know Your Data

1. Open notebook LABS 06: Know Your Data
2. Follow the instructure

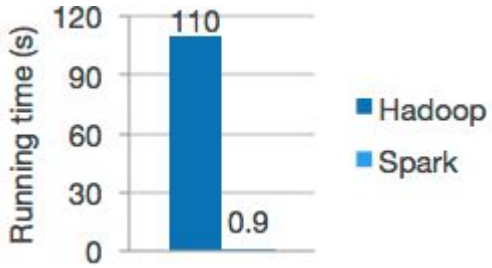
CODING TIME

CHAPTER 07

Answering Business Question

What is Apache Spark?

- Lightning-fast unified analytics engine for large-scale data processing
- The main feature of Spark is its in-memory cluster computing that increases the processing speed of an application
- Run workloads 100x faster



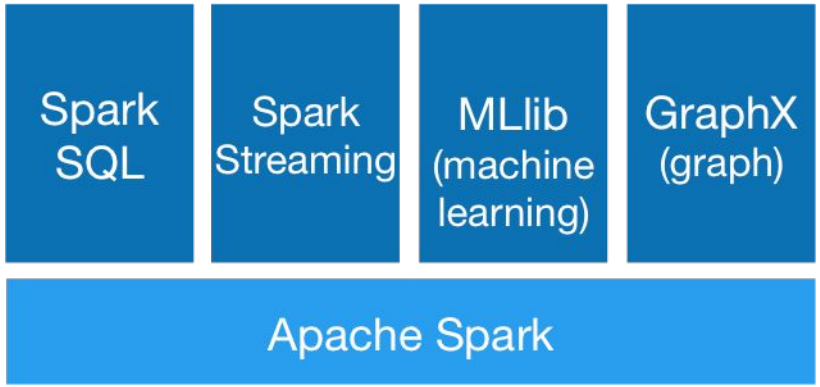
Logistic regression in Hadoop and Spark

- Spark is one of Hadoop's sub project developed in 2009 in UC Berkeley's AMPLab by Matei Zaharia
- Apache Spark is built by a wide set of developers from over 300 companies.
- Since 2009, more than 1200 developers have contributed to Spark

Why Apache Spark?

- Its fast
- Integrate with Hadoop and its ecosystem and can read existing data
- Provide high level API : Scala, Java, Python, R
- Most of machine learning algorithm are iterative
- Can be implemented in standalone mode, Amazon EC2, Mesos and YARN

Component



1. **Spark Core**
   Provide core component for in memory distributed data processing

2. **SparkSQL**
   Spark SQL lets you query structured data inside Spark programs, using either SQL or a familiar DataFrame API. Usable in Java, Scala, Python and R

3. **Spark Streaming**
   Spark Streaming brings Apache Spark's language-integrated API to stream processing, letting you write streaming jobs the same way you write batch jobs

4. **MLlib**
   Provide scalable machine learning library and high-quality algorithms, 100x faster than MapReduce

5. **GraphX**
   a library added in Spark 0.9 that provides an API for manipulating graphs
   (e.g., a social network's friend graph)

Let Use Python

Why Python

- ➔ It’s a lot simpler, and this is just an overview
- ➔ Don’t need to compile anything, deal with JAR’s, dependencies, etc

But

- ➔ Spark itself is written in Scala
- ➔ Scala’s functional programming model is a good fit for distributed processing
- ➔ Give fast performance (Scala compiles to Java bytecode)
- ➔ Less code & boilerplate stuff thanJava
- ➔ Python is slow

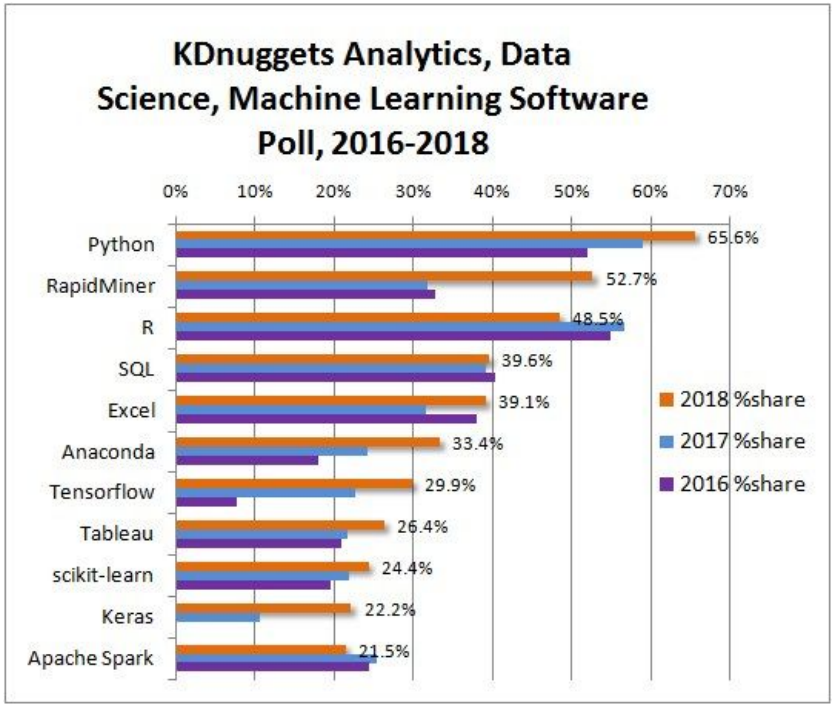
Let Use Python

- Python code

```
nums = sc.parallelize([1, 2, 3, 4])
squared = nums.map(lambda x: x * x).collect()
```

- Scala

```
val nums = sc.parallelize(List(1, 2, 3, 4))
Val squared = nums.map(x => x * x).collect()
```



LABS 07: Answering Business Question

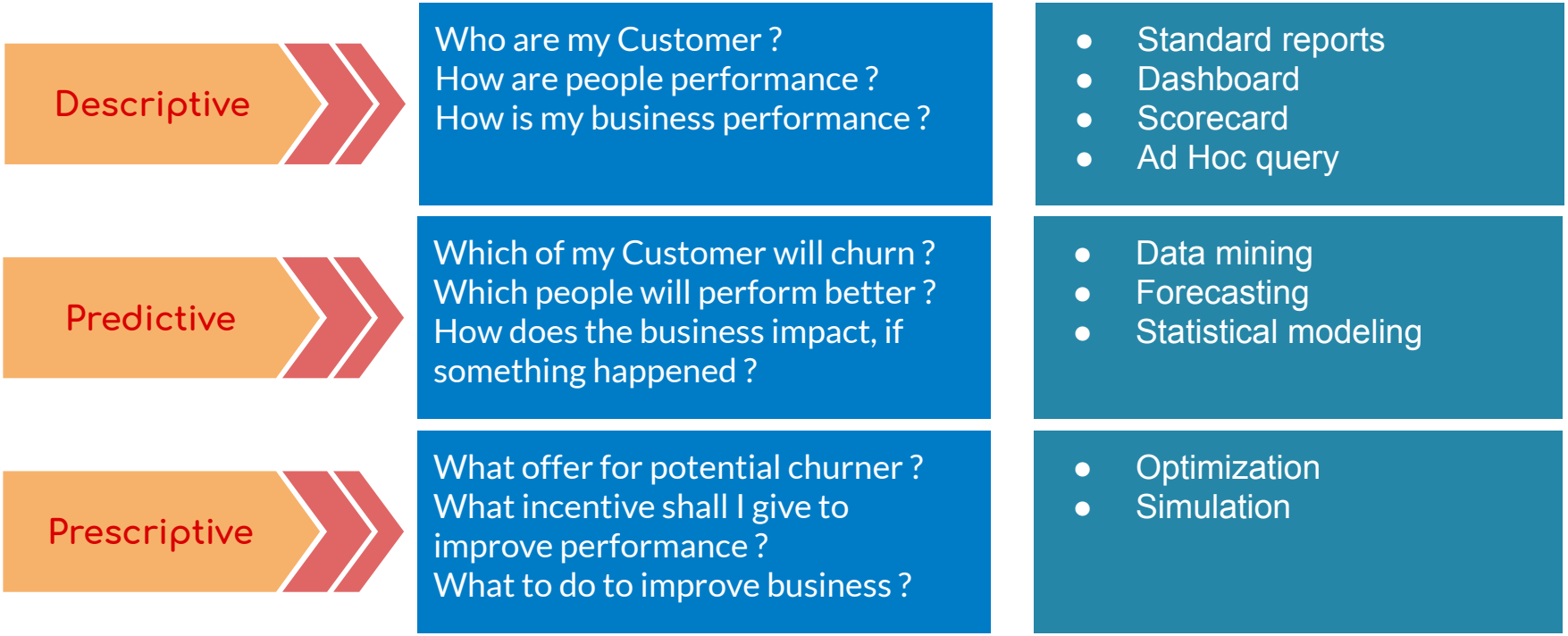
1. Open notebook LABS 07: Answering Business Question
2. Follow the instructure

CODING TIME

CHAPTER 08

Artificial Intelligent

Data Analytics Method



What’s Next?

Artificial Intelligent

LABS 08: Machine Learning With Spark

1. Open notebook LABS 08: Machine Learning With Spark
2. Follow the instructure

CODING TIME

Deep Learning

Artificial Intelligent

the science of getting computers to act in specific ways without explicitly programming them to do so.

Machine Learning

the science of getting computers to act in specific ways without explicitly programming them to do so.
There are a number of machine learning methods or algorithms that can be applied to almost any data problem, i.e

- Regression
- K-Means
- Decision tree
- Random Forest
- etc

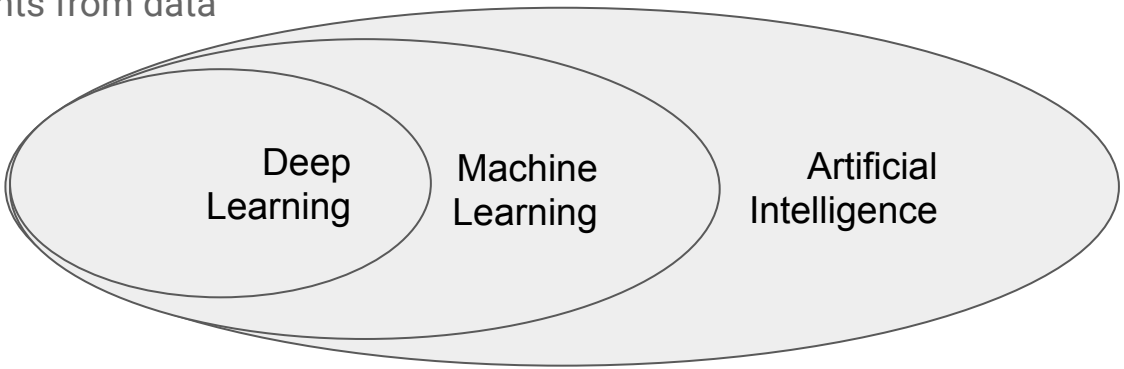
Deep Learning

a form of machine learning that is inspired by the structure of the human brain and is particularly effective in feature detection.
Open source framework :

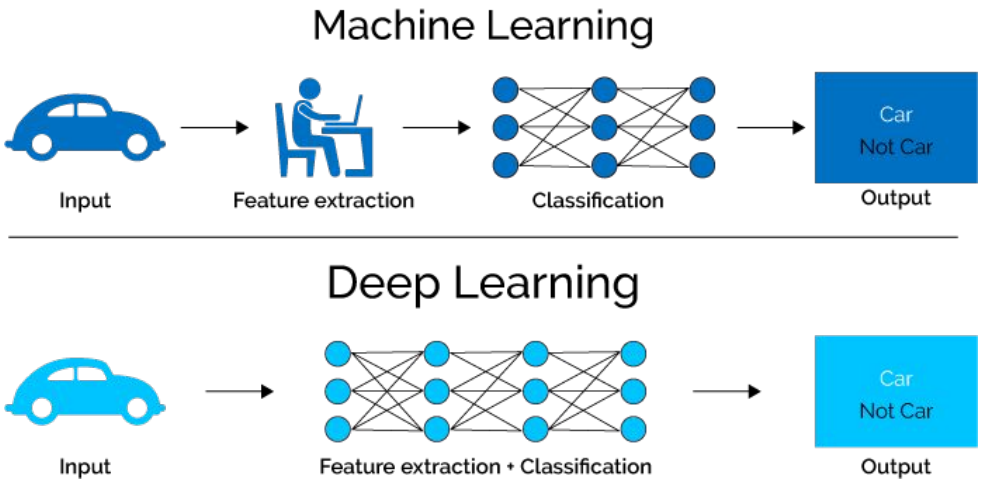
- Theano
- Tensor Flow
- Torch
- Caffe
- MXNet
- etc

What is Deep Learning?

- Machine learning algorithms based on learning **multiple levels** (i.e deep) **of representation/ abstraction** (1)
- Learning algorithms derive meaning out of data by using a hierarchy of **multiple layers** of units (neurons)
- Each neuron/node computes a weighted sum of its inputs and the weighted sum is passed through a nonlinear function, each layer transforms input data in more and more abstract representations
- Learning = find optimal weights from data



Deep Learning vs Classical Machine Learning



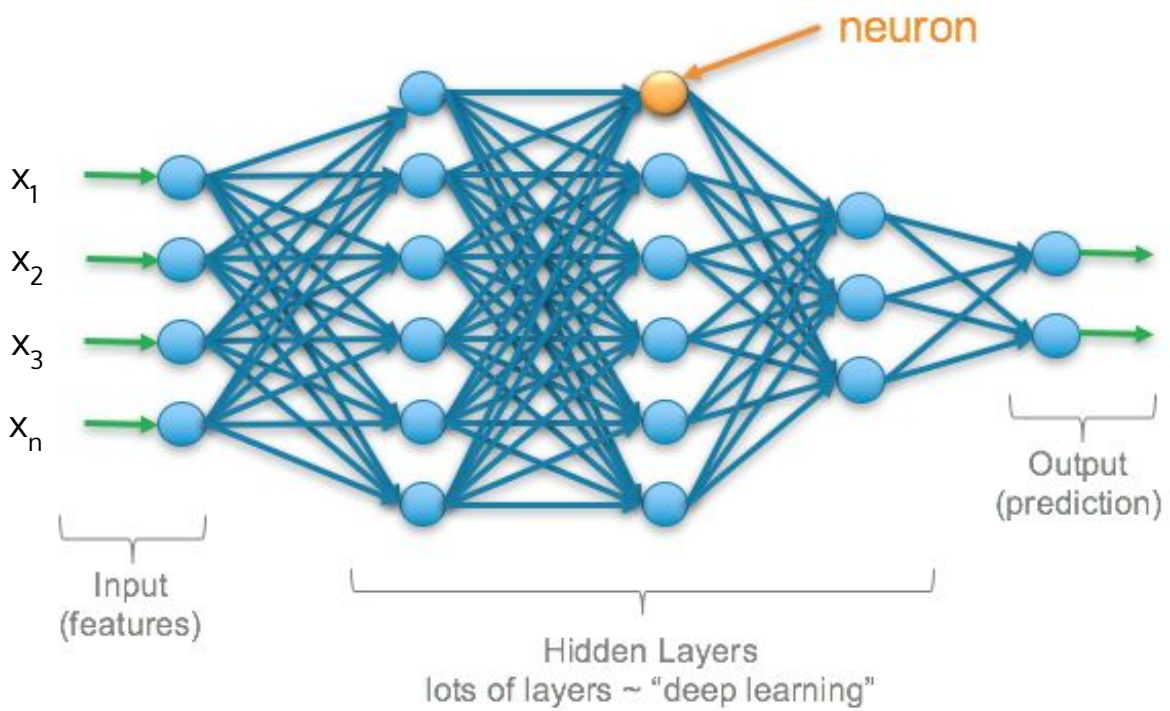
- In classical machine learning, most of the features used require identification of domain experts
- Deep networks scale much better with more data than classical ML algorithms
- Deep learning techniques can be adapted to different domains and applications far more easily than classical ML

Why Now?

- Exponential data growth (and the ability to Process Structured & Unstructured data)
- Faster & open distributed systems (Hadoop, Spark, TensorFlow, …)
- Faster machines and multicore CPU/GPUs
- New and better models, algorithms, ideas:
  - Better, more flexible learning of intermediate representations
  - Effective end-to-end joint system learning
  - Effective learning methods for using contexts and transferring between tasks

"The analogy to deep learning is that the rocket engine is the deep learning models and the fuel is the huge amounts of data we can feed to these algorithms." - Andrew Ng

Multilayer Perceptron



$y_1 = P(cat|x)$

$y_2 = P(dog|x)$

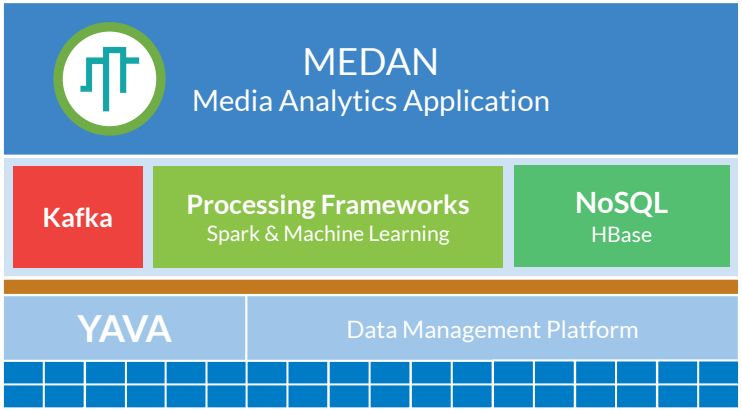
GAN Implementation

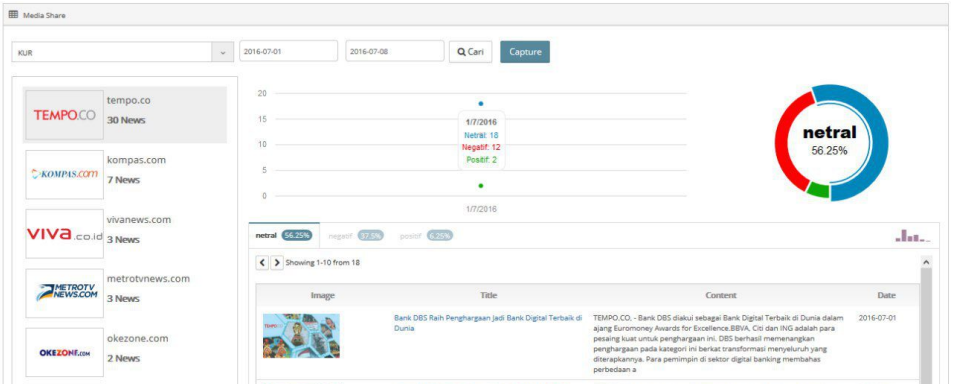
Figure 5: 1024 × 1024 images generated using the CELEBA-HQ dataset. See Appendix F for a larger set of results, and the accompanying video for latent space interpolations.

Paper :
Progressive Growing Of GANS For Improved Quality, Stability, And Variation - NVIDIA

https://arxiv.org/pdf/1710.10196.pdf

Media Analytics

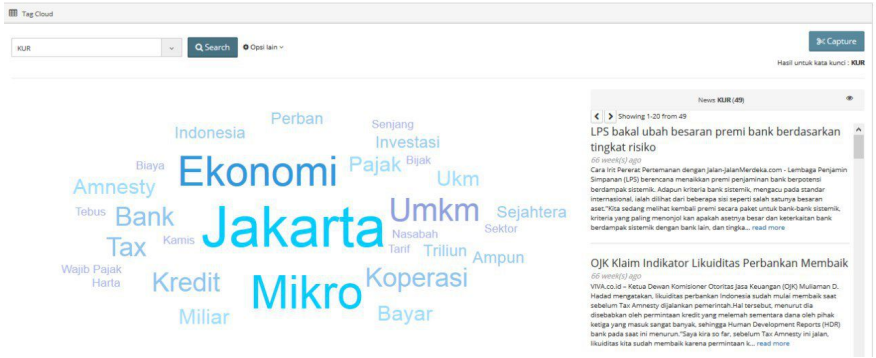


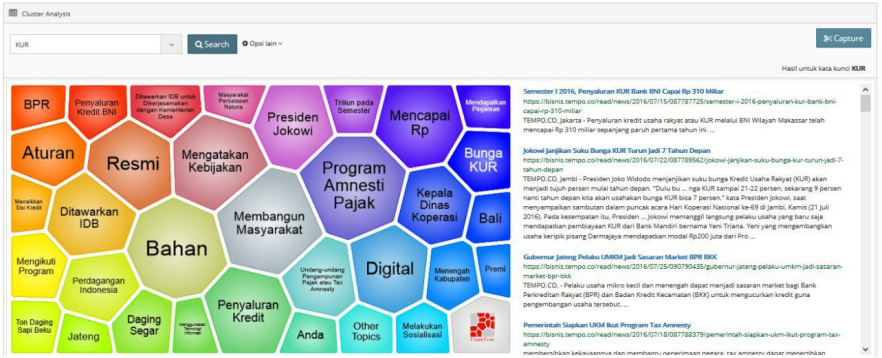


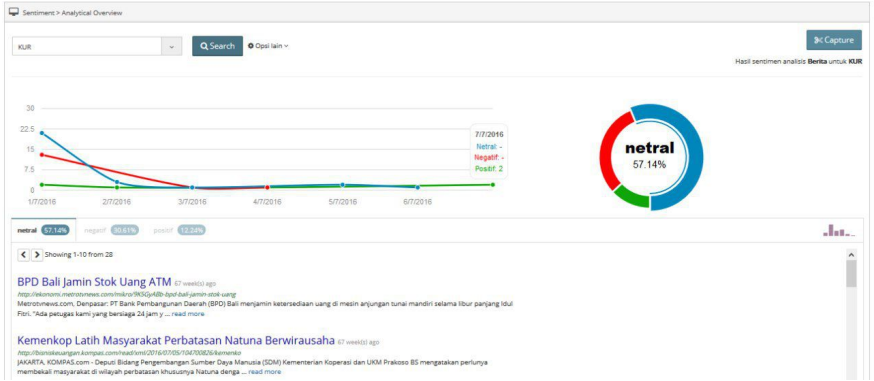
Capabilities :

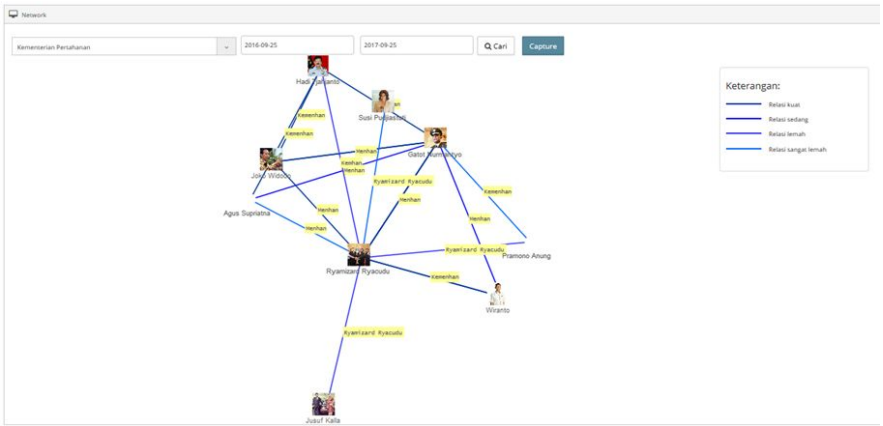
- Crawling Media Online, Forums, SocMed
- Sentiment Analytics, Tag CLouds, Cluster Analytics, Geolocation, Network Analytics, Top Person, etc

Media Analytics

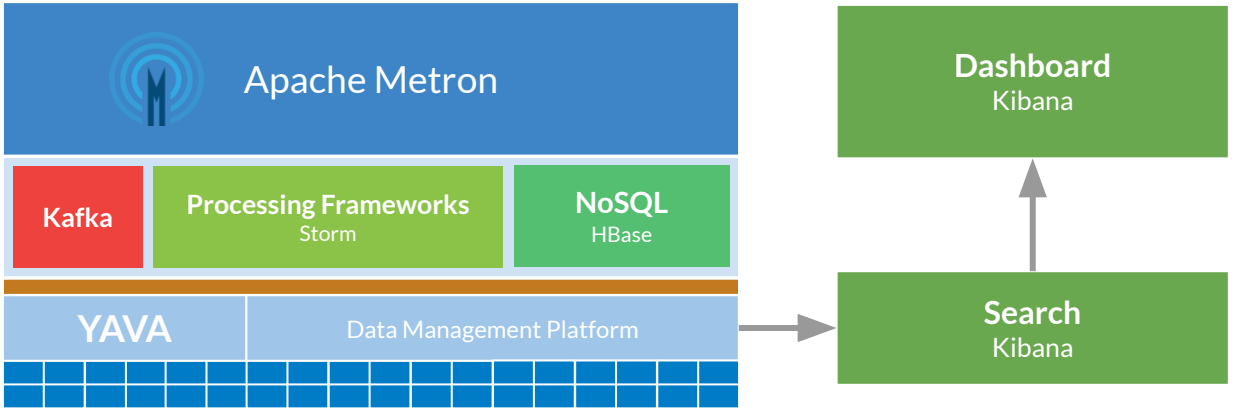






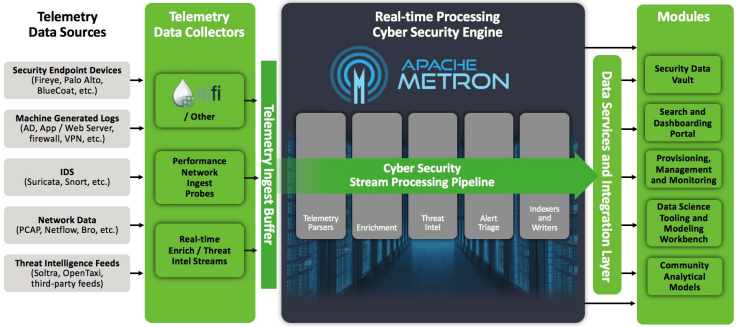


Network Security Analyzer

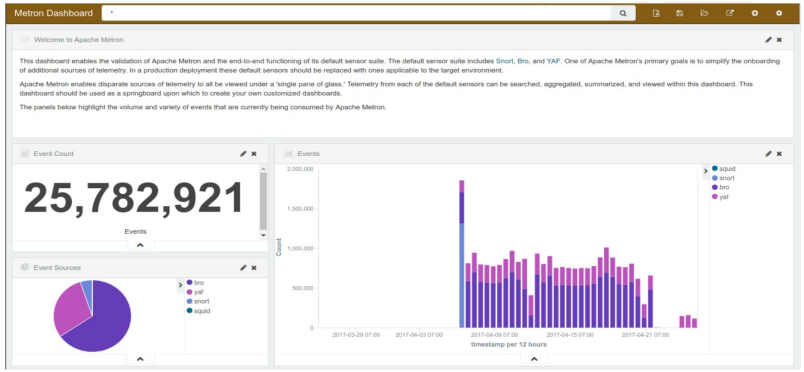


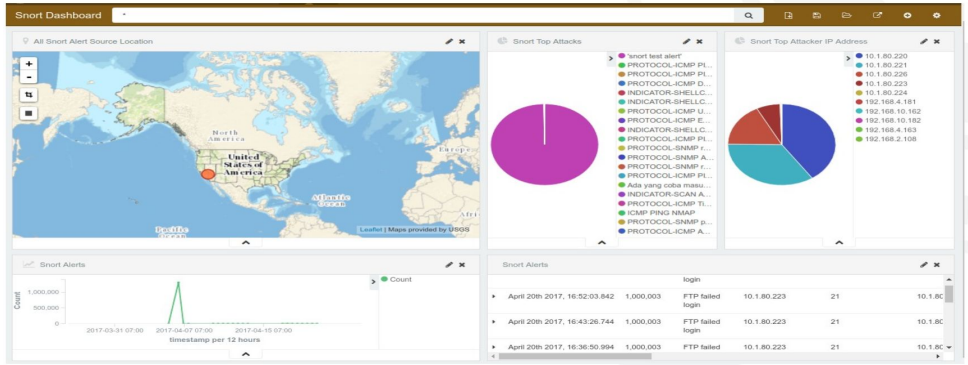
Capabilities :

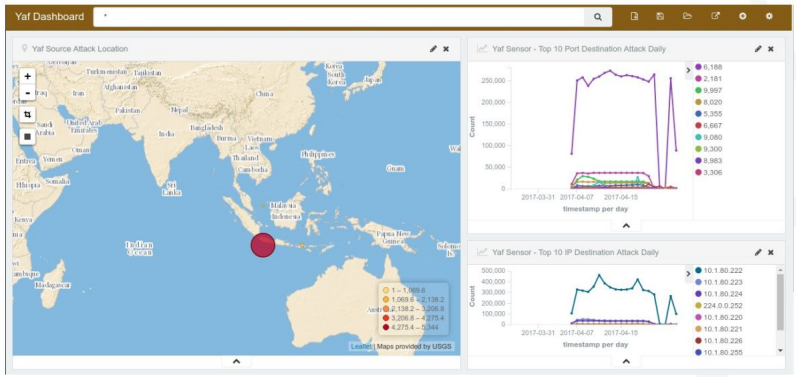
- Security Data Lake
- Pluggable Framework
- Threat Detection Platform
- Incident Response Application

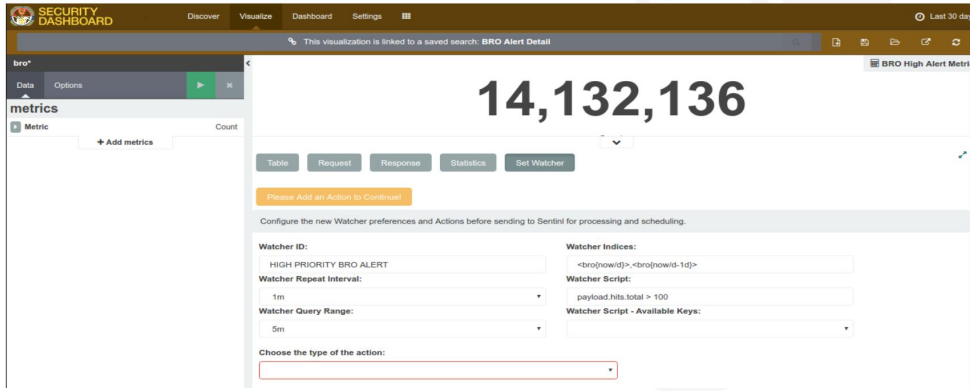


Network Security Analyzer

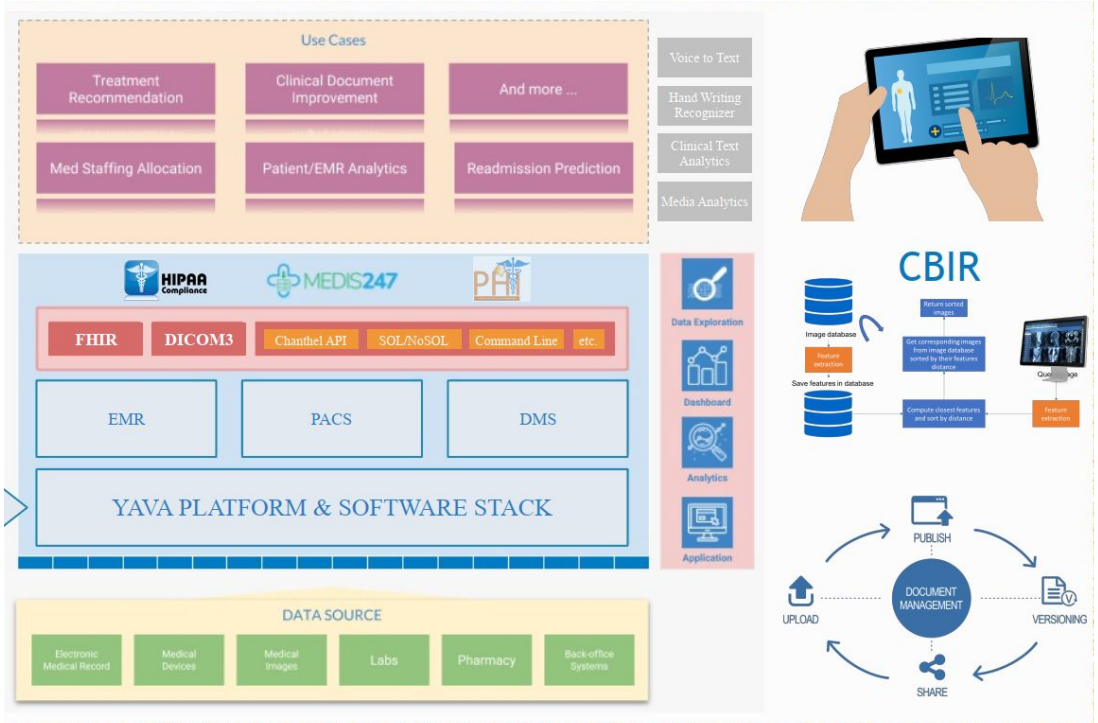








Medis247



- Analytics Dashboard
- Electronic Medical Records - HL7
- Integration with ICU Devices
- Automatic Speech Recognition

In the Past

Jumat 07 April 2006, 12:44 WIB

Demo Sopir Taksi Menolak Blue Bird Diwarnai Bogem Mentah

- detikNews

Senin 10 Desember 2007, 10:51 WIB

Seribuan Sopir Taksi Semarang Kembali Demo Tolak Blue Bird

- detikNews

In The Future

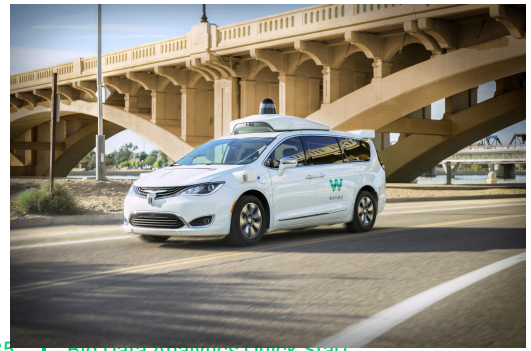
Waymo launches its first commercial self-driving car service

Waymo One's on-demand autonomous rides come with human backup for now.

Jon Fingas, @jonfingas
12.05.18 in Transportation

6
Comments

422
Shares



The rumor was true: Waymo's self-driving car service is here. The company has launched Waymo One, its first commercial ride hailing offering. People in part of the metro Phoenix area can use an app to ask for an autonomous vehicle 24/7 much like they would ridesharing cars, complete with price estimates and trip reviews. Up to three adults and a child can travel at once. To no one's surprise, though, Waymo is starting cautiously -- it's hoping to avoid further collisions and ease the community into a driverless world.


The True
WINNER
Local Appliance
Local Content, Local Support
+
Local Consulting
Local Certification, Local Engineer
=
National Benefit
Prosperity