BIG DATA & MACHINE LEARNING FOR IMPATIENT

Ramakrishna Addanki
Tech Lead – Data Engineering & ML

Ramakrishna Addanki

Tech Lead – Bigdata & Machine Learning

Tech Lead – Big Data & Machine Learning @ Atos Nederland B.V

I am currently working with AIRFRANCE KLM from Atos, leading data engineering and Machine learning tracks for 3 teams.

Specialized in Data Engineering and Data Lakes.

Hortonworks Certified Professional Big Data Engineer.
Microsoft Azure Cloud Certified Big Data Engineer.
Coursera Certified Google Professional Data Engineer.
Oracle Certified Professional.

M.C.A from GITAM University, B.C.A from Nagarjuna University.

Connect with me on

https://www.linkedin.com/in/ramakrishna-addanki-36a788a2

ramakrishnaaddanki.slack.com

Clients worked with

AIRFRANCE KLM GROUP

Nokia Siemens Networks

RENOM Renewing Our Future

SIEMENS

vodafone

ETIQUETTE

SLIENCE YOUR PHONE

ASK QUESTIONS

AGENDA OF SEMINAR

1) The Buzz word “DATA”

2) Buzz word “DATA” in “Industry”

3) Datafication turned into bigdata.

4) Classification of Bigdata

5) Sources of Data

6) Data Modelling

7) Data File Formats

8) Big Data Problem

WHAT DO YOU KNOW ABOUT BIGDATA & ML ?

Using Mentimeter

BUZZ WORD “DATA”

It is an “iPhone”.
It is white in color.
It is designed by Apple.
It is iPhone 8.

Dimensions: 138.4 x 67.3 x 7.3 mm (5.45 x 2.65 x 0.29 in)
Display size: 4.7 inches, 60.9 cm^2

It weighs 148 g (5.22 oz)

consists of a round, hollow puri (a deep-fried crisp crepe),
tamarind chutney, chili, chaat masala,
potato, onion or chickpeas.

It is called as Panipuri.

(Facts && Descriptions) = Data

BUZZ WORD "DATA" "IN" "INDUSTRY"

a) Activity Data

b) Conversational Data

c) Photo and Video Image Data

d) Sensor Data

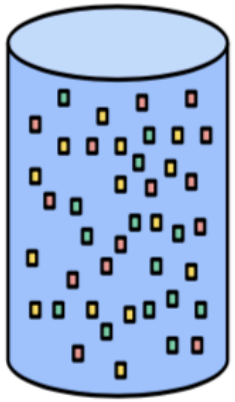
e) Internet of Things Data

BIG DATA CLASSIFICATION

Data can be in two variants, which is as follows:

Data at Rest

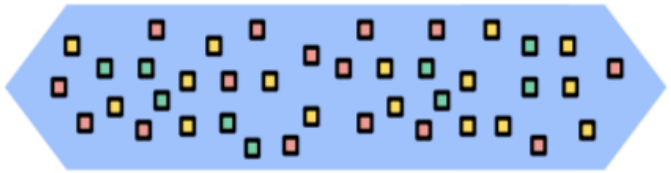
Bounded



Finite data set
Is complete regardless of time
Typically at rest in a common durable store

Data in motion

Unbounded



Infinite data set
Is never complete, especially when considering time
Stored in multiple temporary, yet durable stores

DATAFICATION [BIG DATA]

How much of data?

Volume

Velocity

How frequent or real time is your data ?

What kind data?

Variety

Veracity

How accurate & applicable is your data to business ?

Will you ever have a PetaByte of data?

A stack of floppy disks higher than twelve empire state buildings

27 years to download over 4G

100 Libraries of Congress

Every tweet ever twittered...50 times

VOLUME

How *small* is a PetaByte?

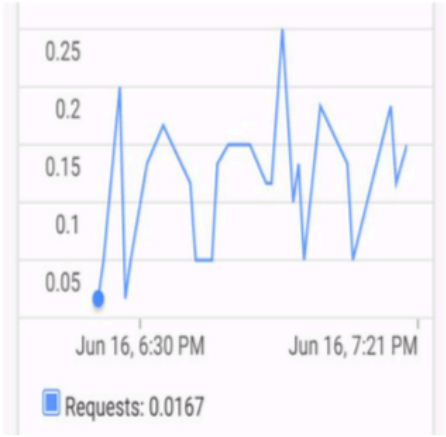
2 micrograms of DNA

1 day's worth of video uploaded to YouTube

200 servers logging at 50 entries per second for 3 years

VARIETY

Traffic sensors along highways

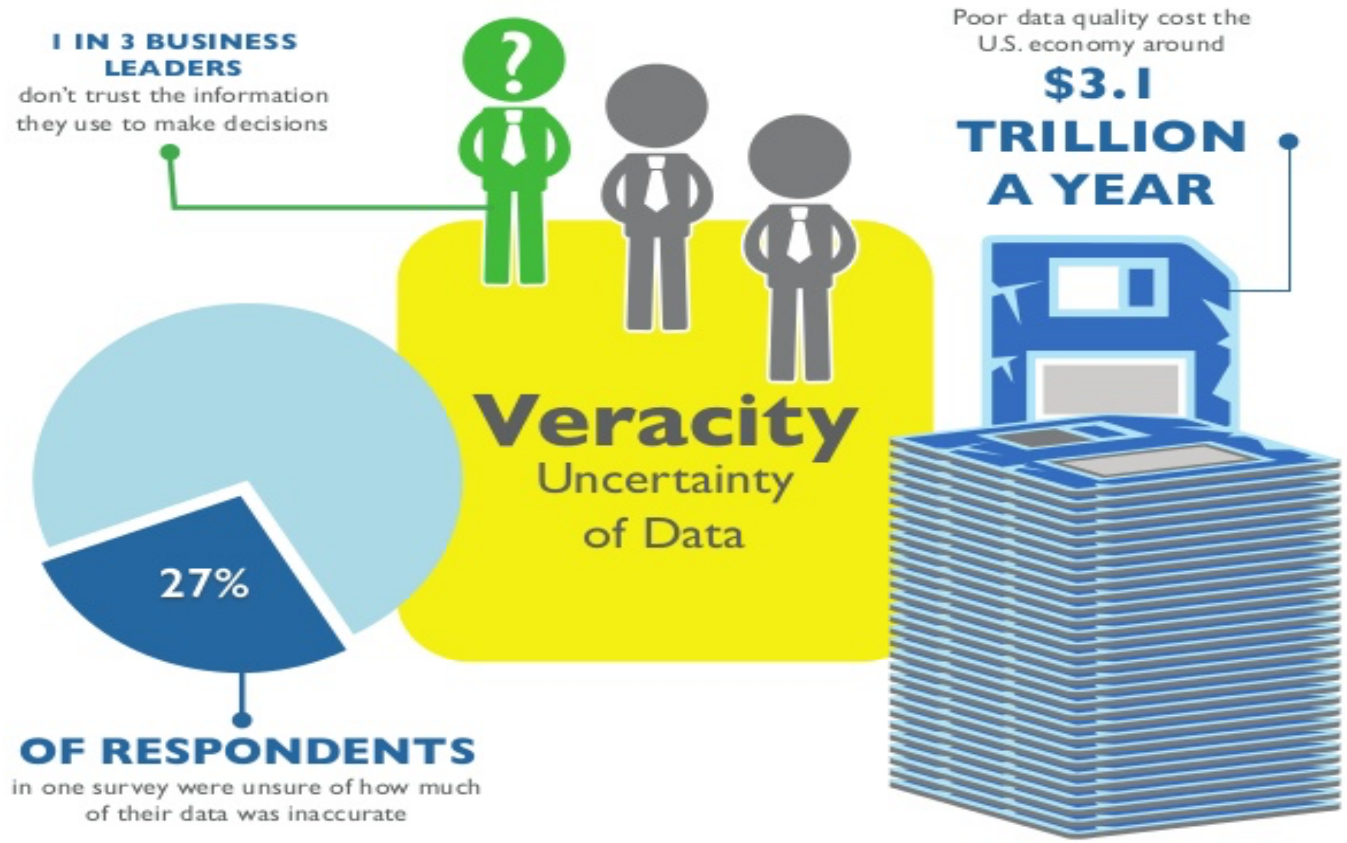


Usage information of Cloud component by every user with a GCP project

Credit card transactions

User moves in multi-user online gaming

VELOCITY


1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality cost the U.S. economy around
$3.1 TRILLION A YEAR
Veracity
Uncertainty of Data
27%
OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate

VERACITY

COMPLETE OVERVIEW [BIG DATA]

How much of data?

- Processing data in batches.
- Perform Map Reduce on massive datasets

How frequent or real time is your data ?

Streaming real time data.
To make immediate insights

What kind data?

Unstructured data
Semi-Structured data
Structured data

How accurate & applicable is your data to business ?

Incompleteness of data to analyze

TURNING BIGDATA INTO VALUE

Datafication of world

a) Activities
b) Conversations
c) Words
d) Voice
e) Social media
f) Browser logs
g) Photos
h) Videos
i) Sensors

Volume

Velocity

Variety

Veracity

Analyzing Bigdata

a) Text Analytics
b) Sentimental Analysis
c) Face Recognition
d) Voice Analytics
e) Movement Analytics
Etc..

Value

SOURCES OF DATA

Why ??????

Data you collect and analyse today

Data you collect but don't analyse

Data you could collect but don't

Data from partners. subsidiaries and 3rd parties

DATA MODELLING

- Structured Data
- Semi-Structured Data
- Unstructured Data

Assume you have a table like this in RDBMS like Oracle or SQL Server

Timestamp	Scheduled Departure Airport	Num_flights
2019-05-22 21:03:00	AMS	4184
2019-05-22 22:01:24	ORD	3610
2019-05-22 05:45:19	DFW	3121
2019-05-23. 07:54:23	AMS	2300
2019-05-22 23:05:09	DEN	2212

Is it possible to analyze & answer the following question based on above data ?

How many total number of flights departed from "AMS" Airport on 22nd and 23rd

YES

Using SELECT & GROUP BY

STRUCTURED DATA

NO

It is not possible directly to run a query on this type of Data.

To run SQL query on this data, you need to ingest this data into DWH (SQL Server / Oracle).

SEMI STRUCTURED DATA

How many total number of flights departed from "CDG" Airport on 22nd and 23rd

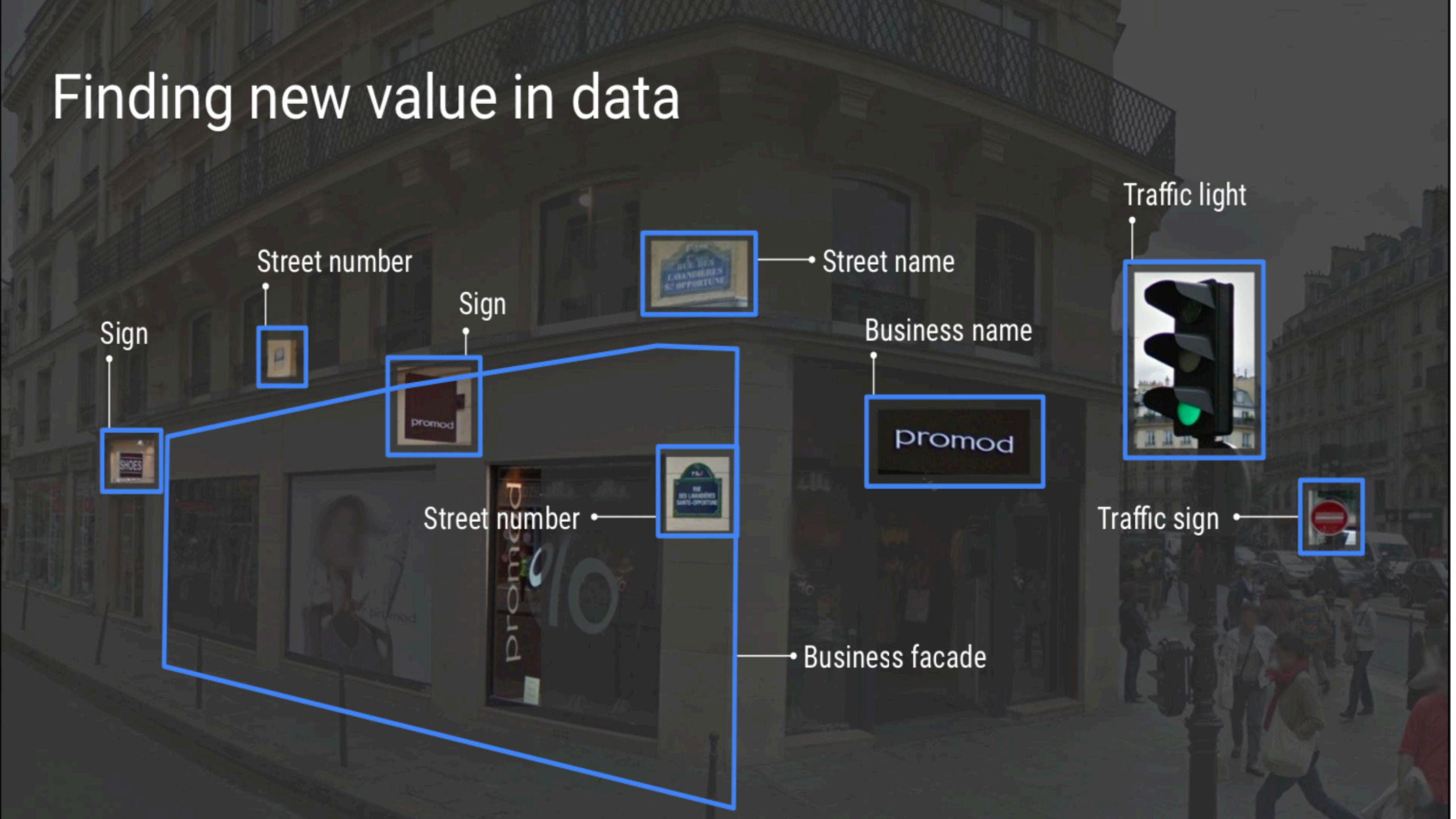
ETL Systems [**E**xtract **T**ransform **L**oad].

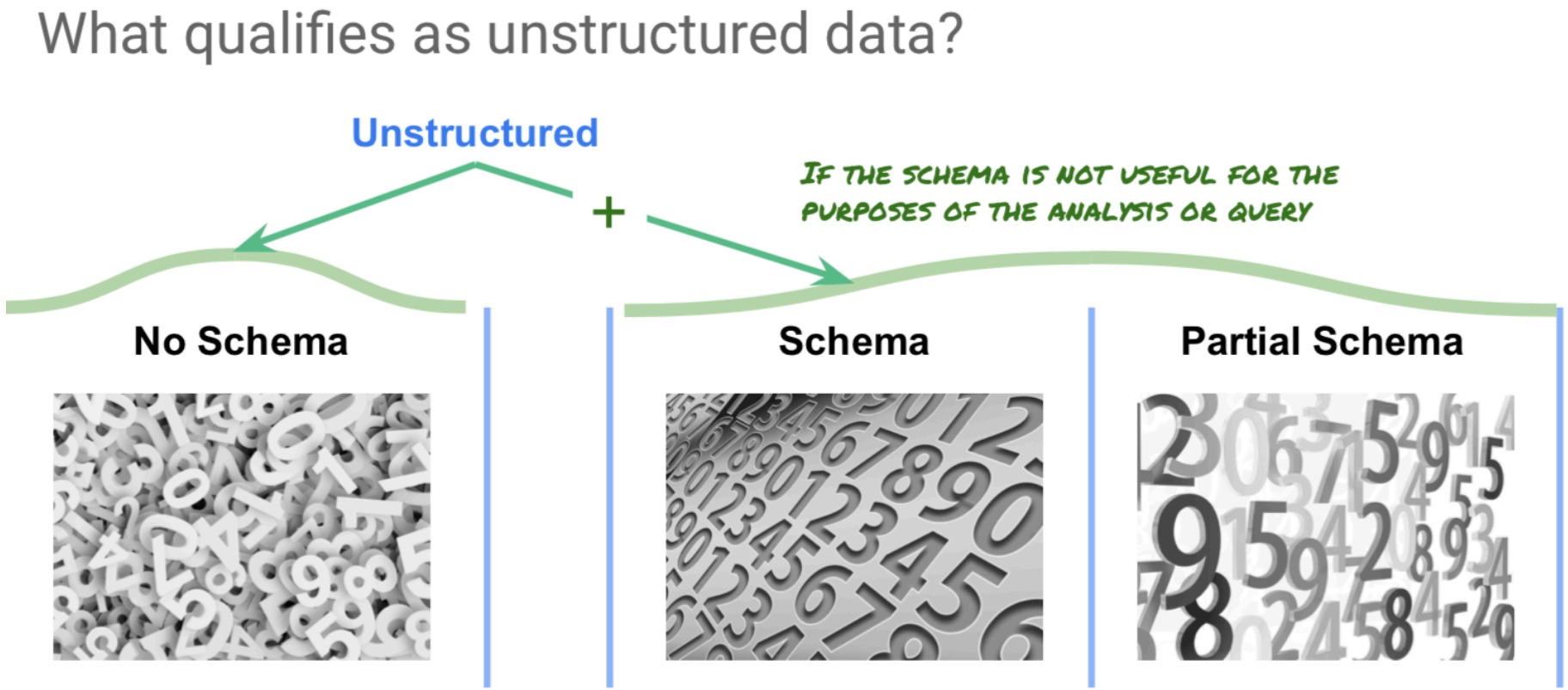

```
[
    {
        Timestamp : 2019-05-22 21:03:00,
        Scheduled_Departure_Airport : "CDG",
        Num_flights : 4184
    },
    {
        Timestamp : 2019-05-22 22:01:24,
        Scheduled_Departure_Airport : "ORD",
        Num_flights : 3610
    },
    {
        Timestamp : 2019-05-22 05:45:19,
        Scheduled_Departure_Airport : "DFW",
        Num_flights : 3121
    },
    {
        Timestamp : 2019-05-23 07:54:23,
        Scheduled_Departure_Airport : "CDG",
        Num_flights : 2300
    },
    {
        Timestamp : 2019-05-22 23:05:09,
        Scheduled_Departure_Airport : "DEN",
        Num_flights : 2212
    }


]
```

UNSTRUCTURED DATA

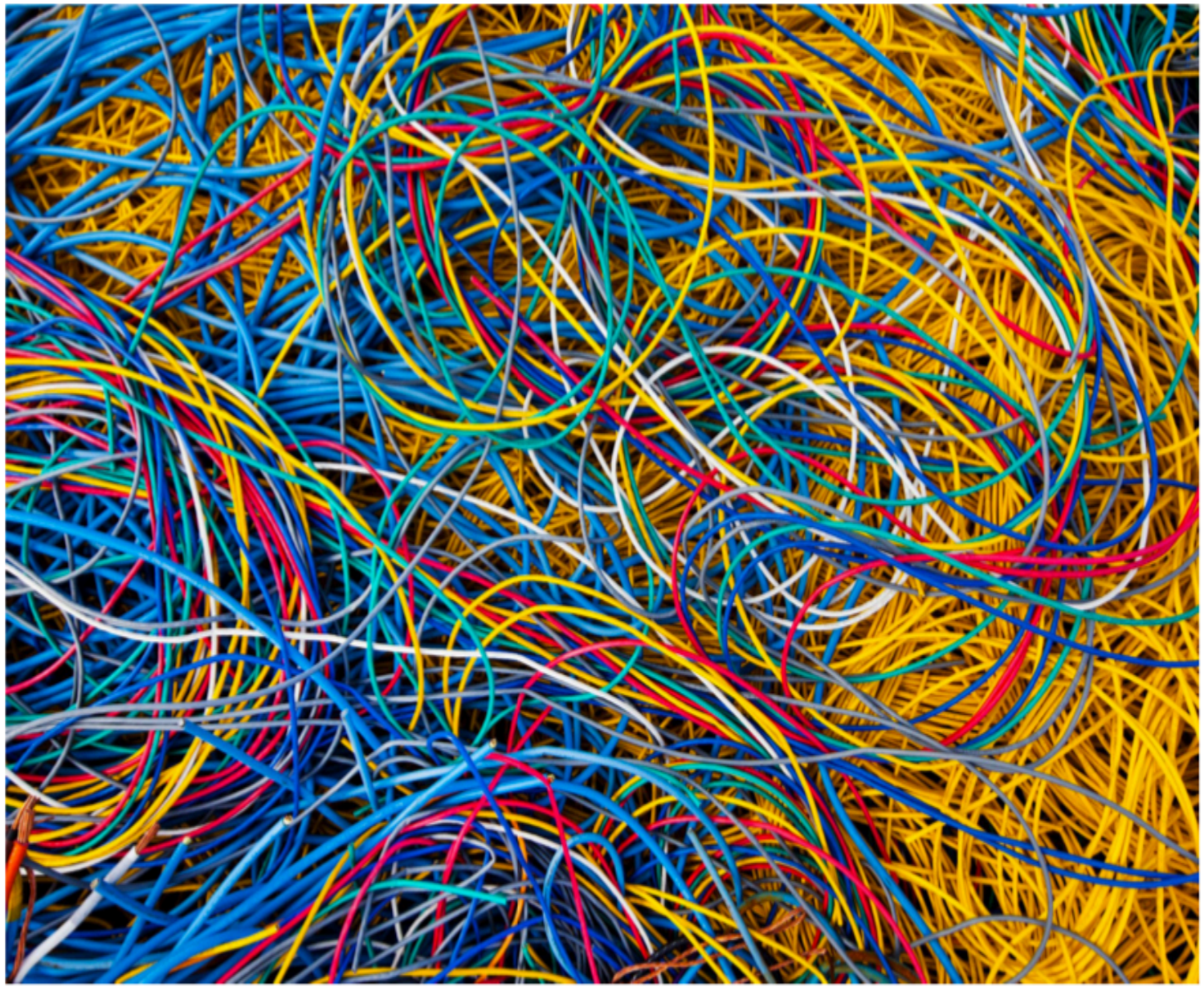
Finding new value in data





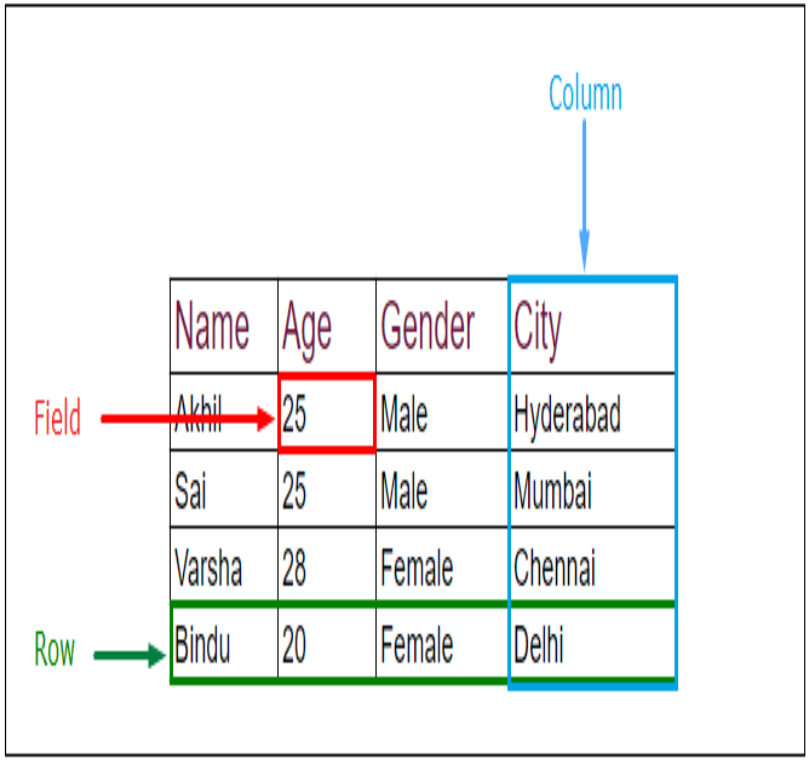
DATA MODELING OVERVIEW

Unstructured data accounts for 90% of enterprise data*

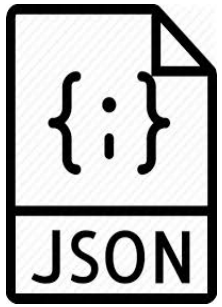


DATA FILE FORMATS

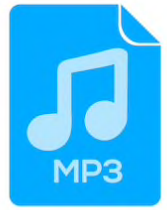
Structured Data : RDBMS Table



Semi Structured file formats:



Unstructured file formats:



BIG DATA FILE FORMATS

These are exclusively designed for big data systems to process data.

Interesting thing about these formats are:

These file formats are Semi-Structured as well as Unstructured.

The variant data types allows this flexibility.

THE BIG DATA PROBLEM

The Problem:

Can we process this all data on single machine or even possible to store this data !

Google

SCALING UP

Solution.

Distribute the data over multiple large machines (aka) cluster.

NEW FRAMEWORK ?

The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung

Google*

&

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

+

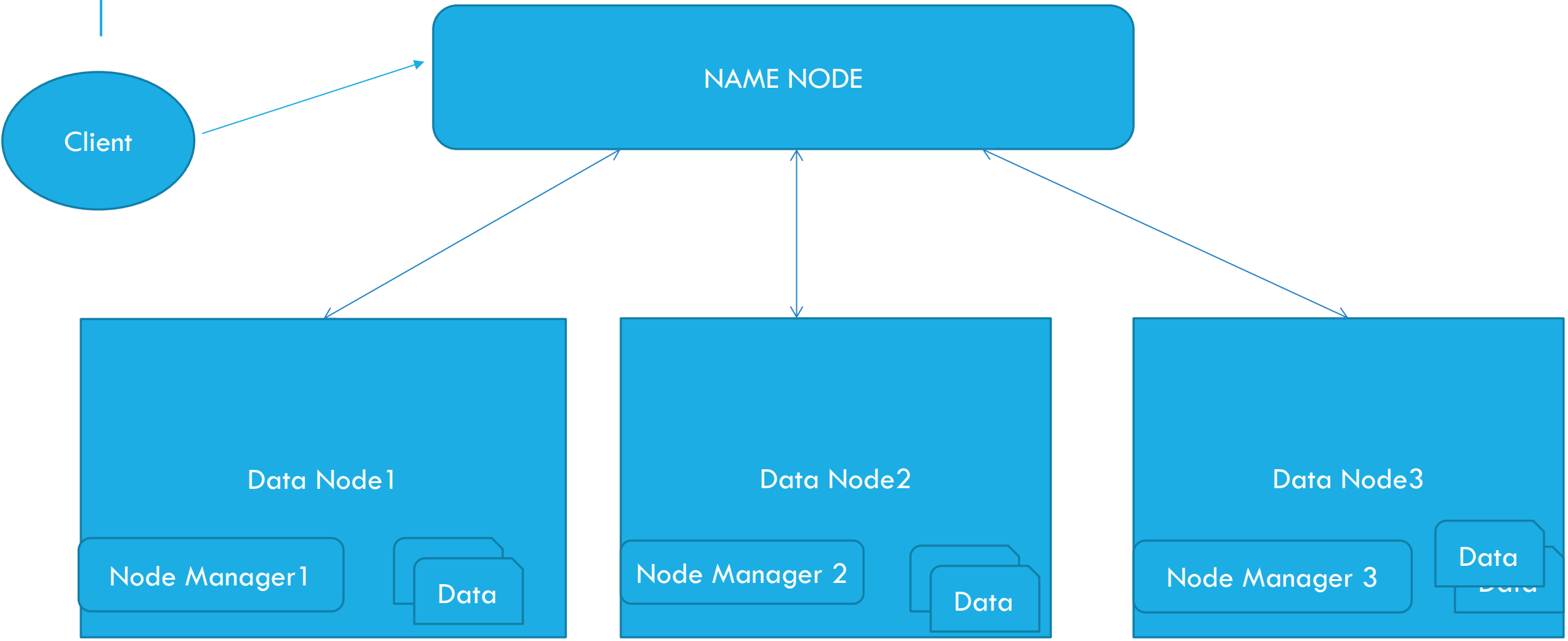
=

Distributed Storage

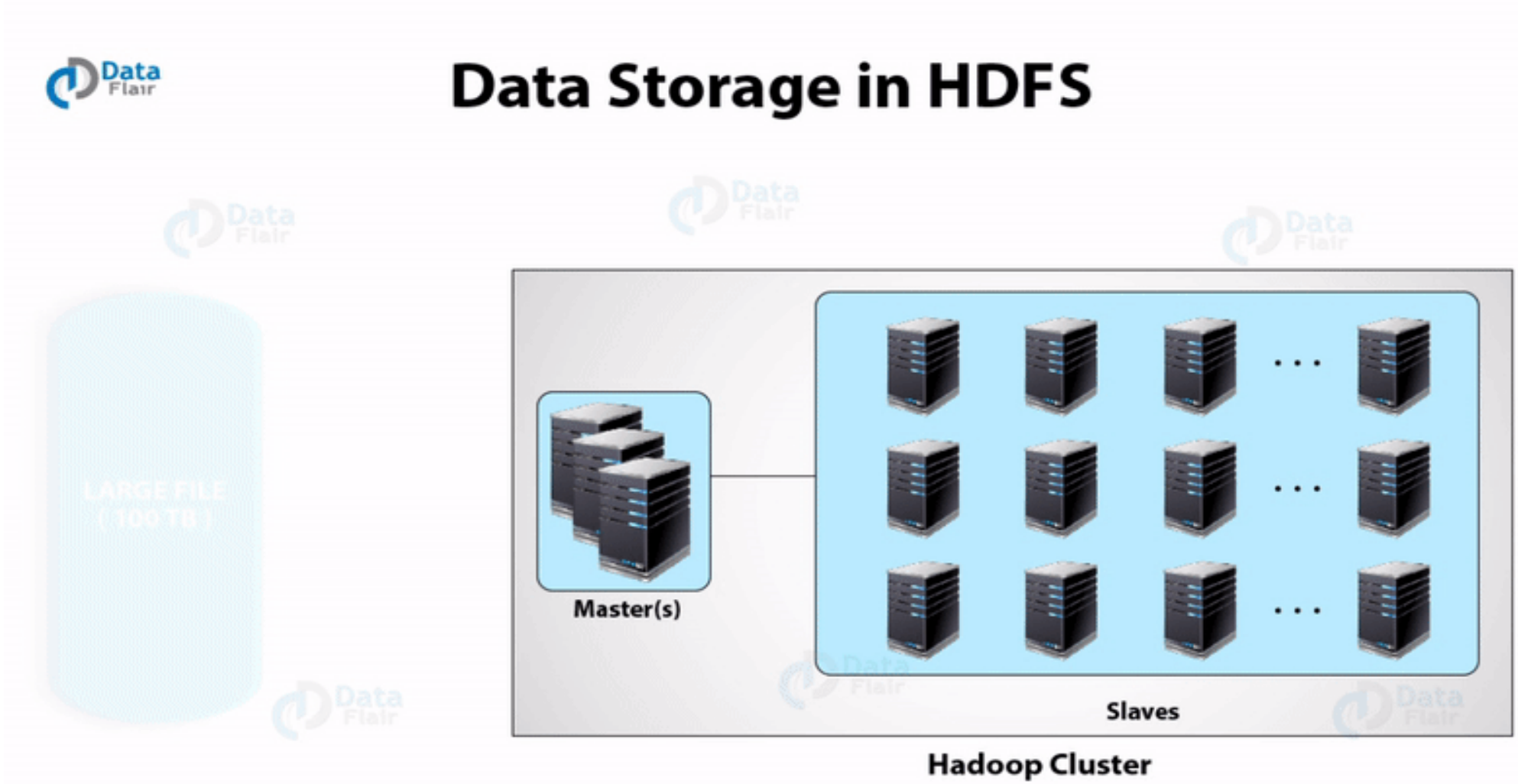
Parallel Processing

GOOGLE BIGDATA DATA CENTERS

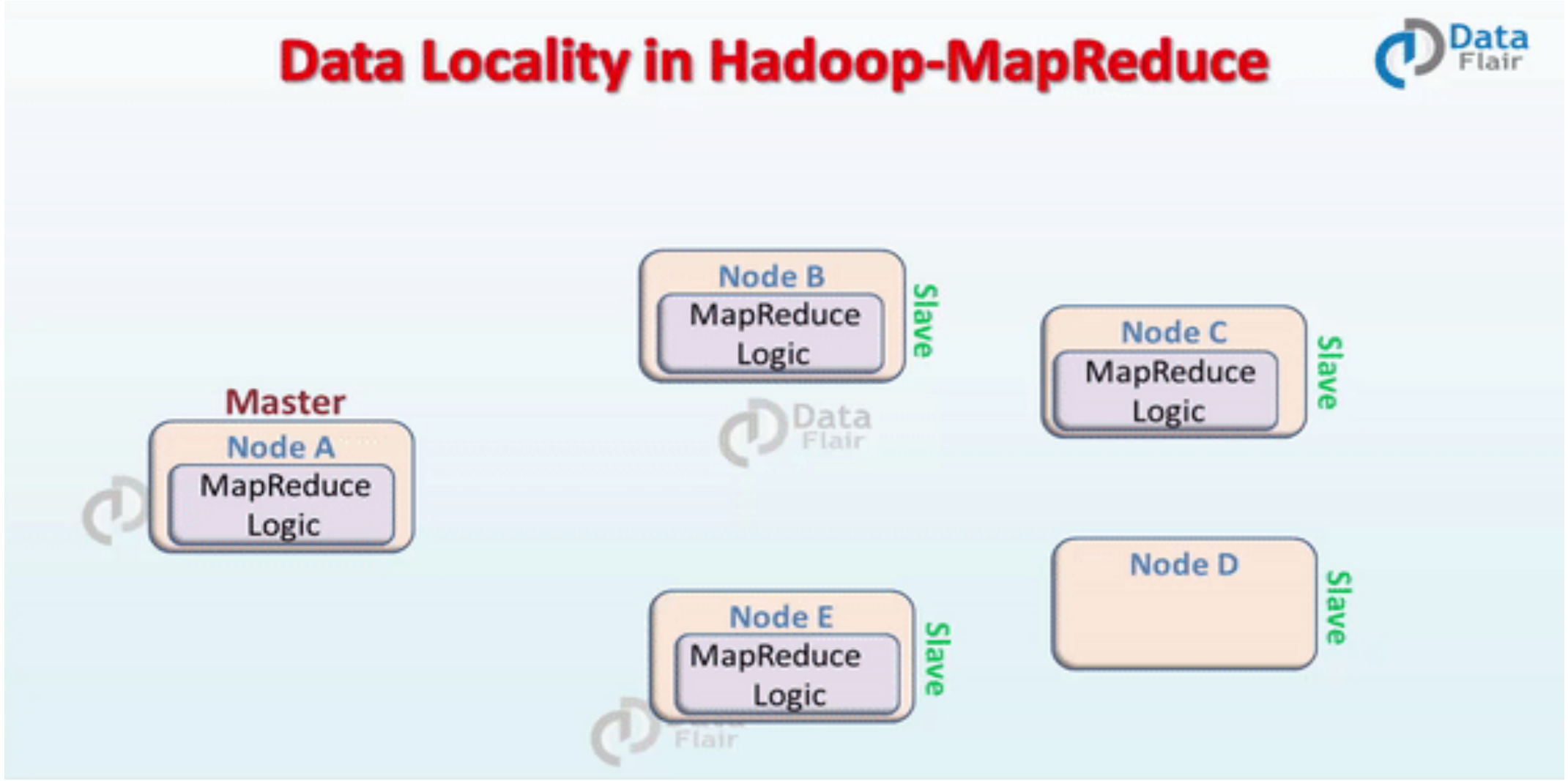
HDFS - COMPONENTS



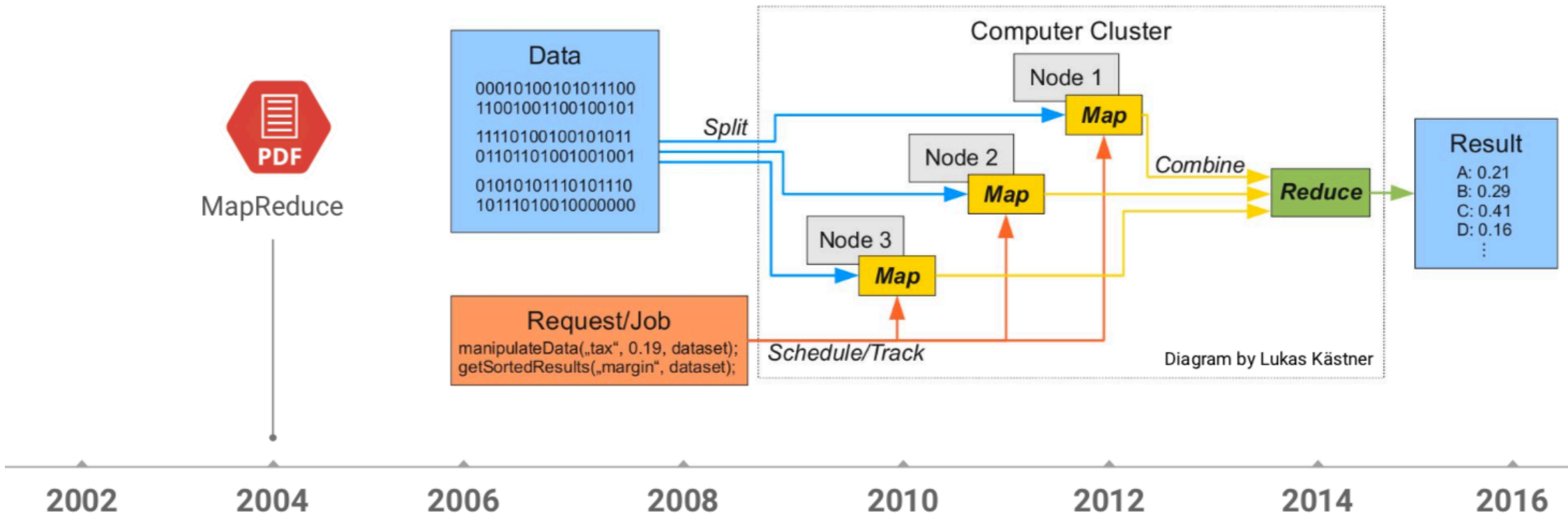
HDFS


Data Flair
Data Storage in HDFS
LARGE FILE
(100 TB)
Master(s)
Slaves
Hadoop Cluster

HDFS – DATA LOCALITY



MapReduce approach splits Big Data so that each compute node processes data local to it



HADOOP DATA PLATFORM

SPARK (In-Memory, Data Flow Engine)

HIVE & DRILL (Analytical SQL-on-Hadoop)

MAHOUT & SPARK MLlib (Machine learning)

PIG (Scripting)

HBASE (NoSQL Database)

ZOOKEEPER & AMBARI (Management & Coordination)

MAPREDUCE (Processing using different languages)

KAFKA & STORM (Streaming)

SOLR & LUCENE (Searching & Indexing)

OOZIE (Scheduling)

DATA LAKE

Resource Management

YARN

Storage

HDFS

Apache Ambari

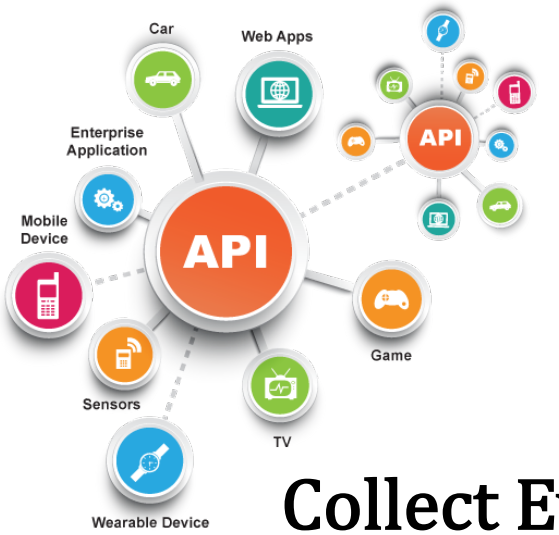
Flume

Unstructured/ Semi-structured Data

Sqoop

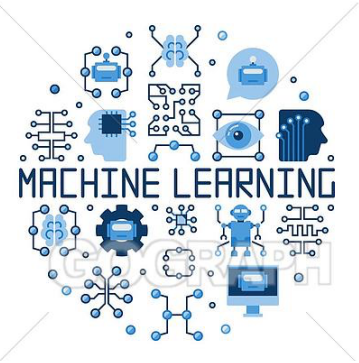
Structured Data

PROMISE OF DATA LAKE



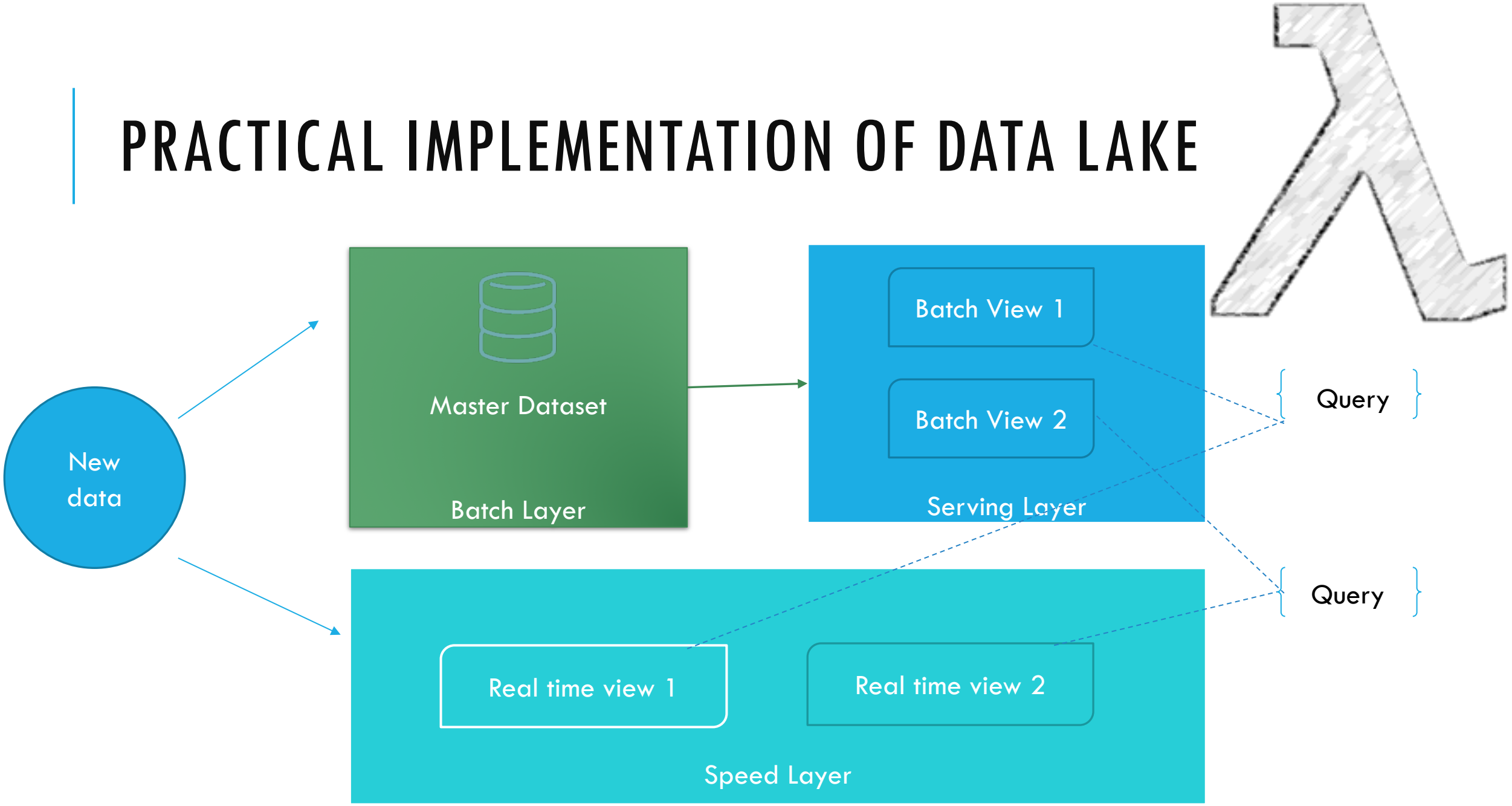
Collect Everything !!

Store it in Data lake

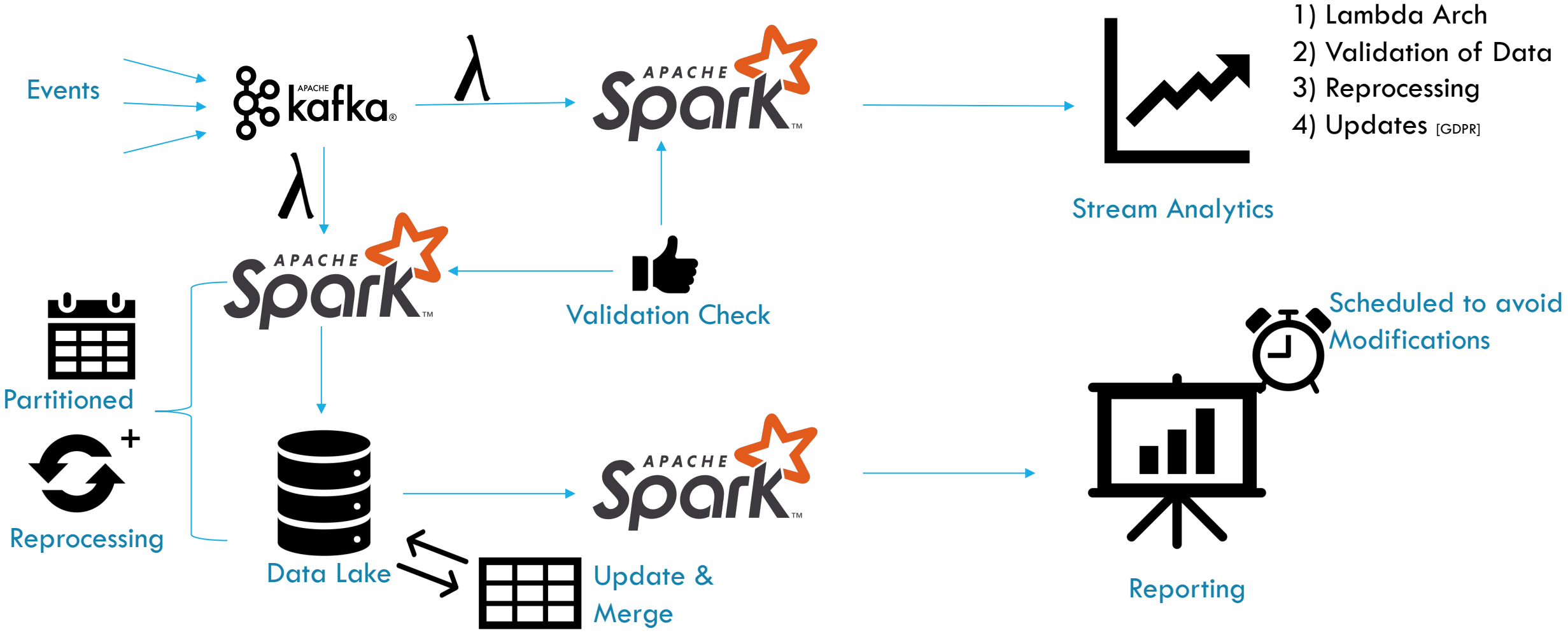


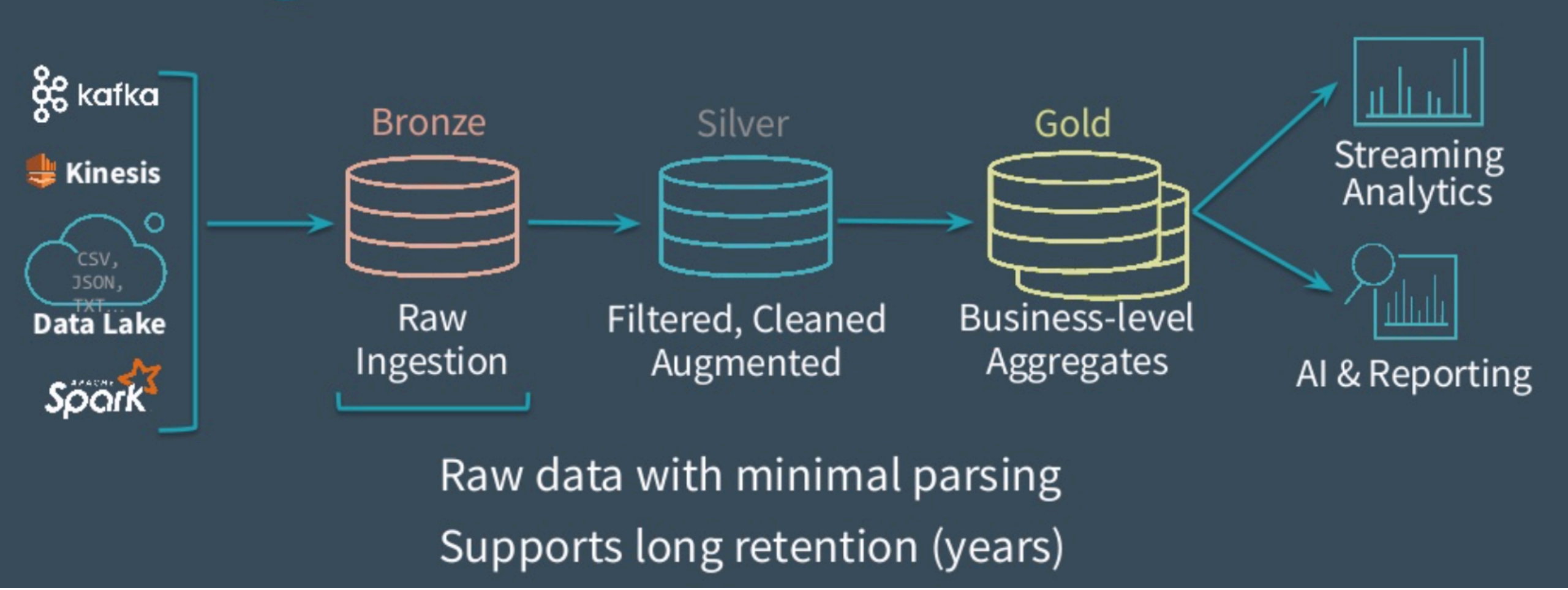
Recommendations
Predictive Maintenance
IoT, Fraud Detections

PRACTICAL IMPLEMENTATION OF DATA LAKE



INDUSTRY APPROACH FOR BIG DATA WORKFLOWS





DELTA ARCHITECTURE

LIVE TWITTER FEED – SENTIMENT ANALYSIS

QUESTIONS ??