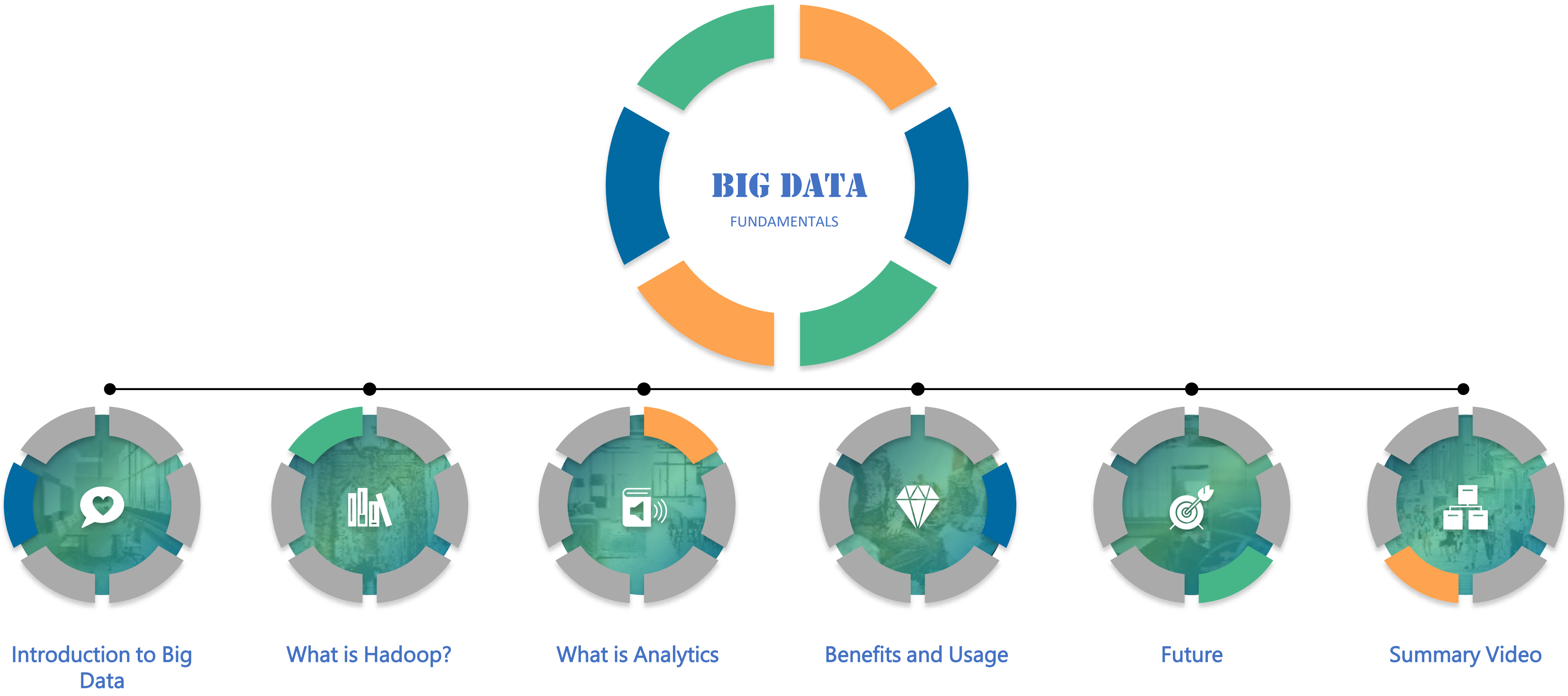
BIG DATA

FUNDAMENTALS

DISCLAIMER: All logos and images used are copyright of respective owners

Contents



BIG DATA

FUNDAMENTALS

Introduction to Big Data

What is Hadoop?

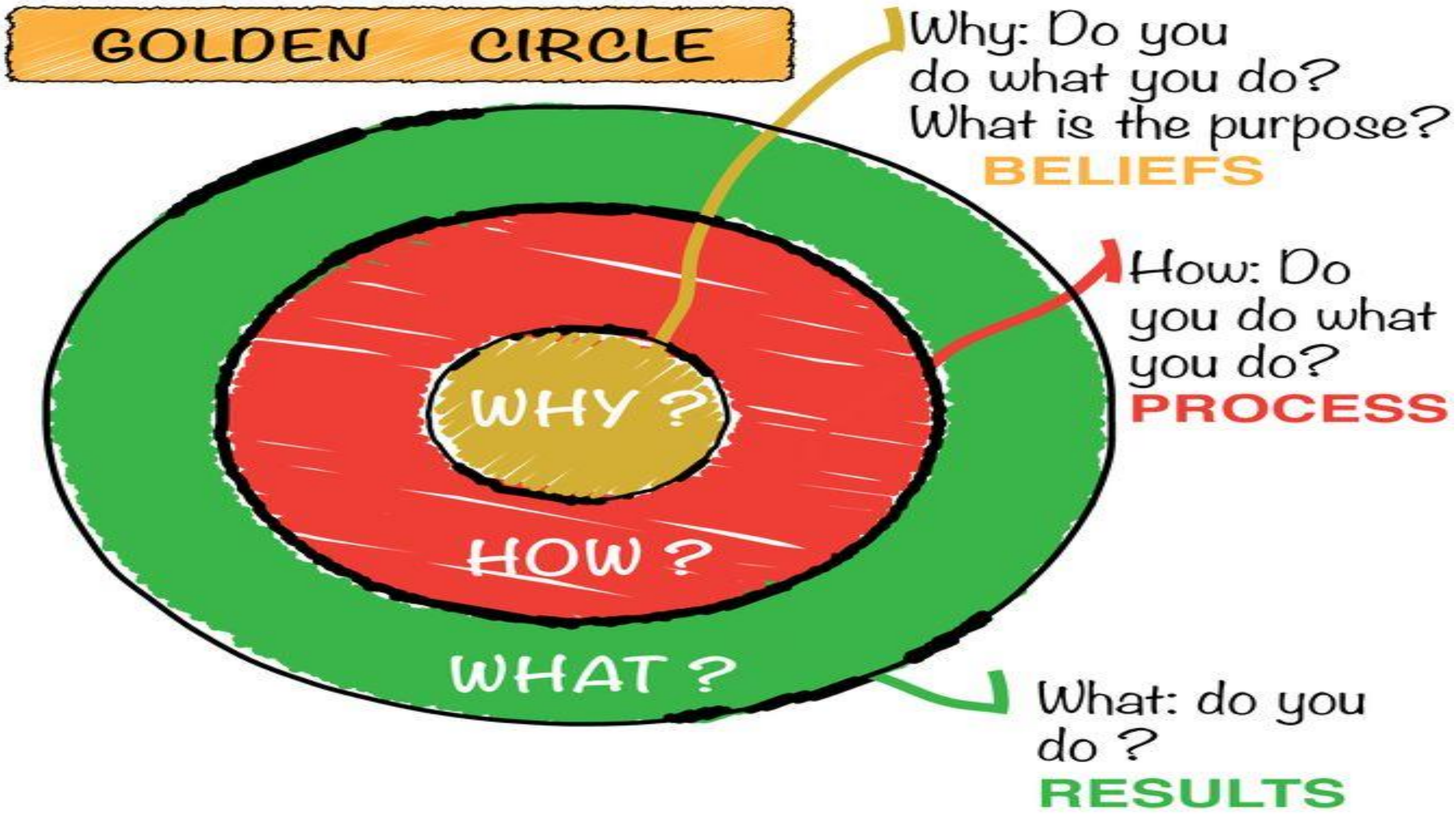
What is Analytics

Benefits and Usage

Future

Summary Video

Introduction to Big Data


GOLDEN CIRCLE
WHY ?
HOW ?
WHAT ?
Why: Do you do what you do? What is the purpose?
BELIEFS
How: Do you do what you do?
PROCESS
What: do you do ?
RESULTS

Introduction to Big Data

WHY?

IT systems over the years

Mainframes

- zOS based systems
- Originally bought computer services to the masses
- Processed Structured data
- CICS based screens
- Analytics via CSV etc format reports

- Data Maturity Level: Few Hundred Gigabyte

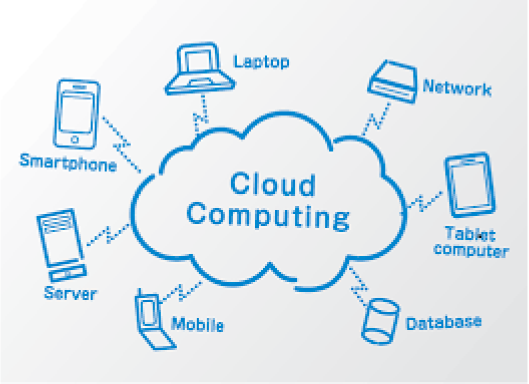
- Descriptive Analytics

Server Based

- Unix based systems
- Introduced GUI
- Processed Structured data
- Batch Analytics via specialized software like Crystal reports etc

- Data Maturity Level: Thousands Terabytes

- Predictive Analytics



Cloud

- Distributed systems
- Accessible to multiple platforms
- Processed Structured\Unstructured data
- Real-time Analytics via specialized software Tableau, Qlikview etc

- Data Maturity Level: Petabyte\Exabyte level

- Prescriptive Analytics

Introduction to Big Data

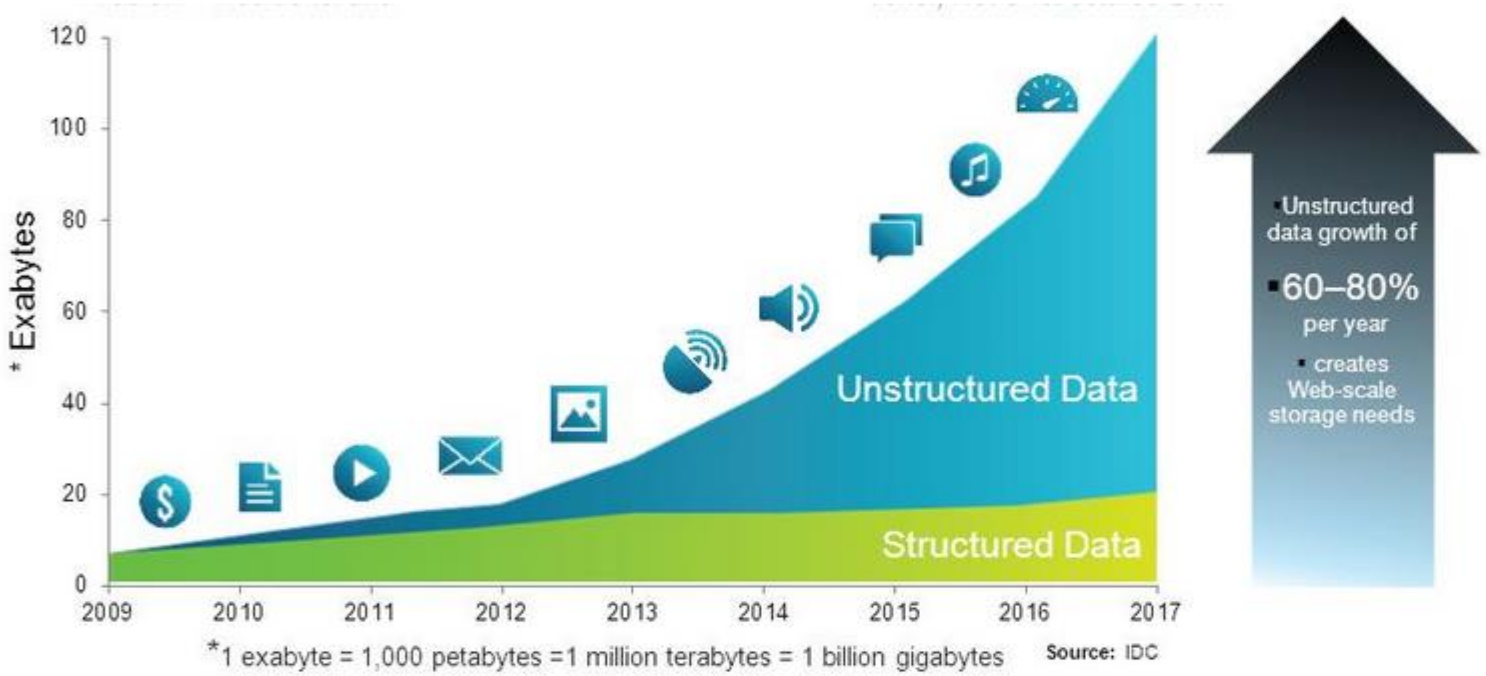
WHY?

Why did need for Big data come into being

As illustrated, data was scaling exponentially, our understanding of data was not. Information is nothing but interpreted data. Though the data increased exponentially, our processing systems did not scale up linearly. Our processing servers grew large till a tipping point.

Houston, we have a problem.

In short, the problem was: More data, less processing power.



*1 exabyte = 1,000 petabytes =1 million terabytes = 1 billion gigabytes

Source: IDC

Moore's Law Is Dead. Now What?

The data volumes are exploding, more data has been created in the past two years than in the entire previous history of the human race.

New Sources like Machine logs, Facebook conversation, Instagram posts, Telegram\WhatsApp messages , IoT devices are expected to contribute to this in future.

Introduction to Big Data

WHAT?

Three main characteristics of Big Data:

❖ Volume
❖ Velocity
❖ Variety

❖ Veracity is newly added.

Big data solutions are not a single elixir, or a silver bullet which will get rid of our data problems.
It is a eco system of several products working together

Example: Uber or Ola Share matching working at real time speeds

Data
from various sources

- Structured
- Semi Structured
- Unstructured

Terabytes of Data is ingested and processed in real time for several applications

Big Data model is :

- Highly scalable
- Flexible
- Reduced Cost
- New products and services-Lesser time to Market*

Introduction to Big Data

HOW?

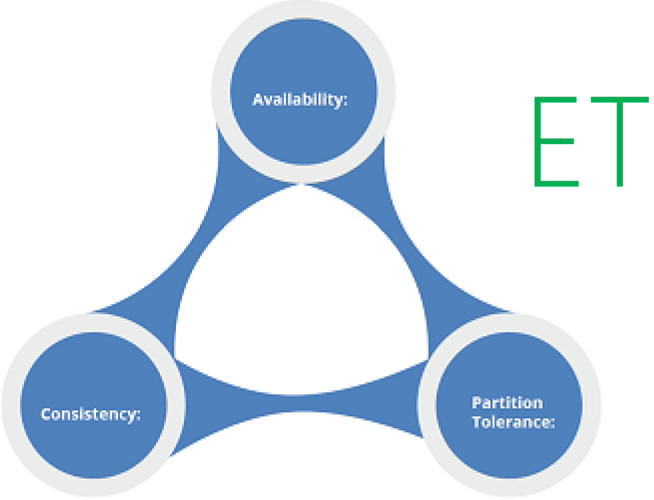
Answer: Distributed computing systems.

Processing Paradigms :

DAG-Spark & Map Reduce

Traditional RDBMS vs NoSQL systems

Visual Guide to NoSQL Systems



ETL Vs ELT

Origins of Modern Big Data

- ❖ Google Bigtable and Google Big Query
- ❖ Yahoo Nutch which became Apache Hadoop

What is Hadoop?

Answer : Apache Hadoop is an open-source software framework which acts as the layer of abstraction which takes care of the traditional data issues and provide infra on which a distributed computing machine can be built

Lets Address the ELEPHANT in the Room

Hadoop Distributions examples:

- Cloudera
- HortonWorks
- MapR
- IBM BigInsight

Myth: Big Data=Hadoop

Truth: Hadoop is just one of the implementations of BD.There are others like Ceph,Disco, Hydra etc

Hadoop remains the most popular with distributions like Cloudera, HortonWorks and MapR

Other Hadoop-related projects at Apache include:

- ❖ Hive
- ❖ Pig
- ❖ Spark
- ❖ Zookeeper

Trademarks and Logos are copyright of respective

What is Hadoop?

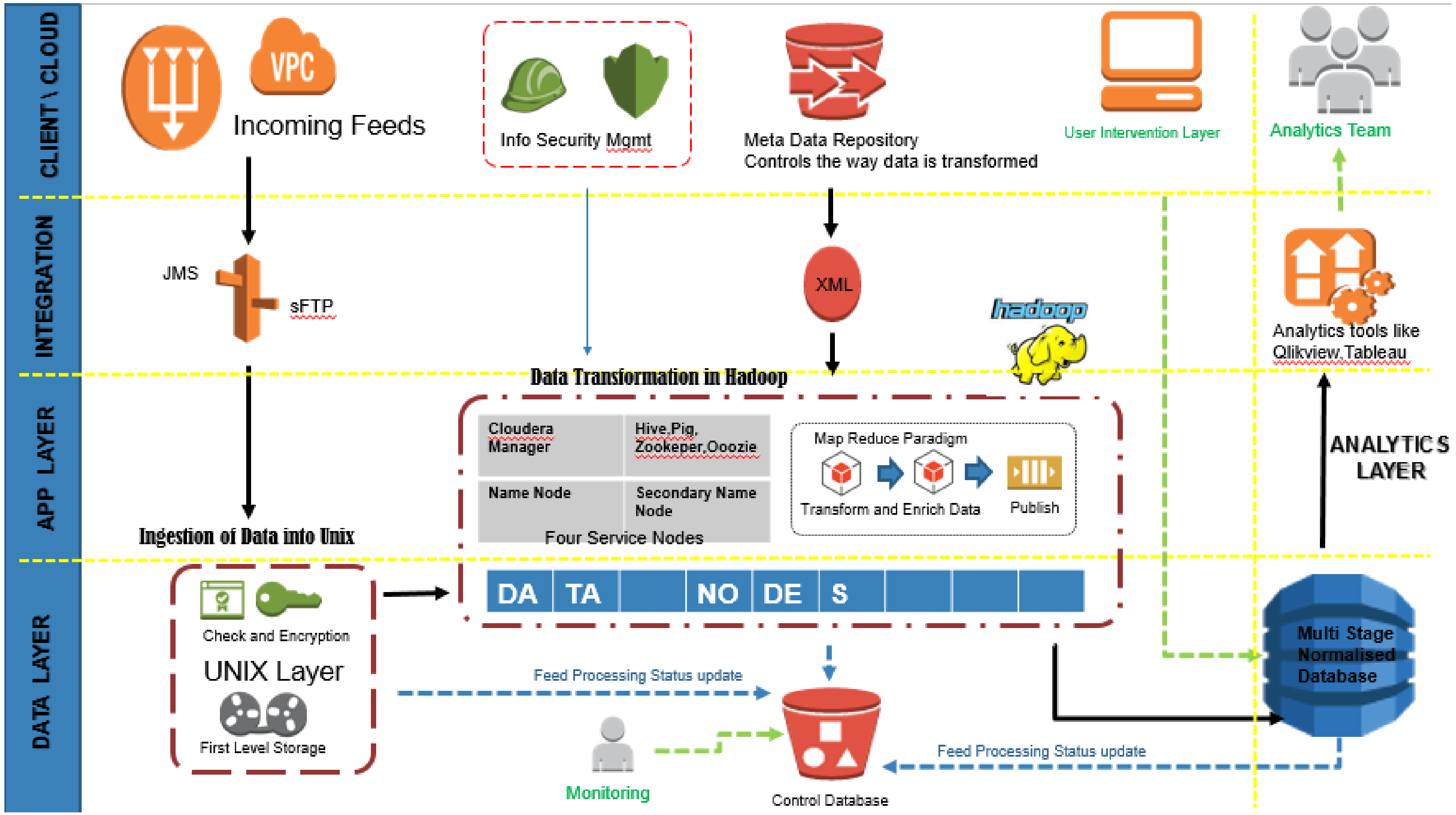
Basic Layers

DATA INGESTION

DATA TRANSFORMATION

DATA ANALYSIS

Sample Architecture: For IT Associates



CLIENT \ CLOUD

INTEGRATION

APP LAYER

DATA LAYER

VPC

Incoming Feeds

Info Security Mgmt

Meta Data Repository
Controls the way data is transformed

User Intervention Layer

Analytics Team

JMS

sFTP

XML

hadoop

Analytics tools like
Qlikview,Tableau

Data Transformation in Hadoop

Cloudera Manager	Hive,Pig, Zookeper,Ooozie
Name Node	Secondary Name Node

Four Service Nodes

Map Reduce Paradigm

Transform and Enrich Data

Publish

ANALYTICS LAYER

Ingestion of Data into Unix

DATA NODES

Check and Encryption

UNIX Layer

First Level Storage

Feed Processing Status update

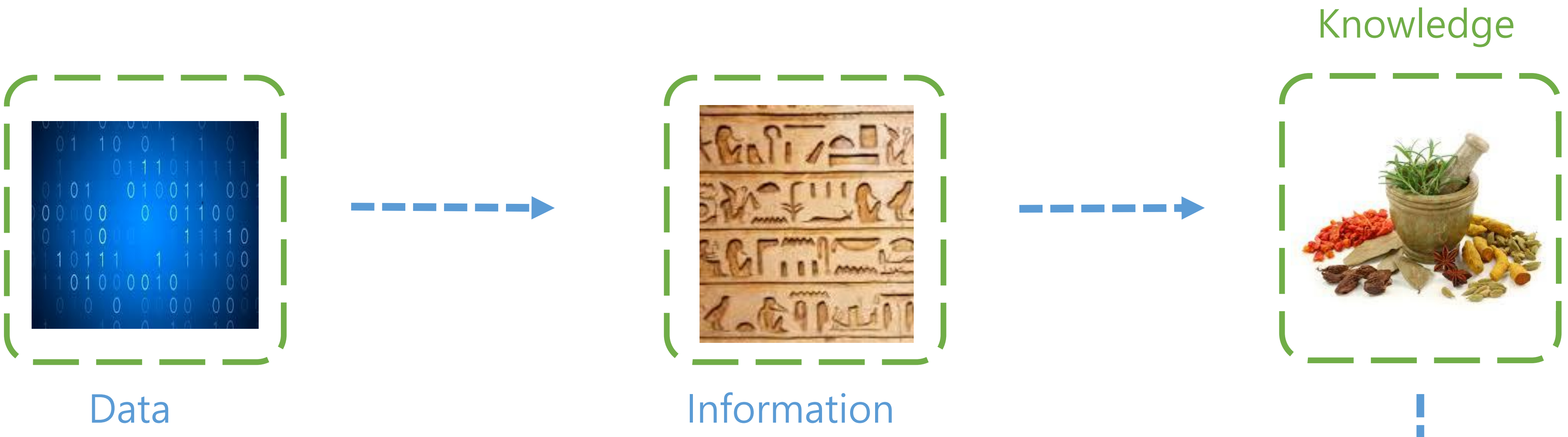
Monitoring

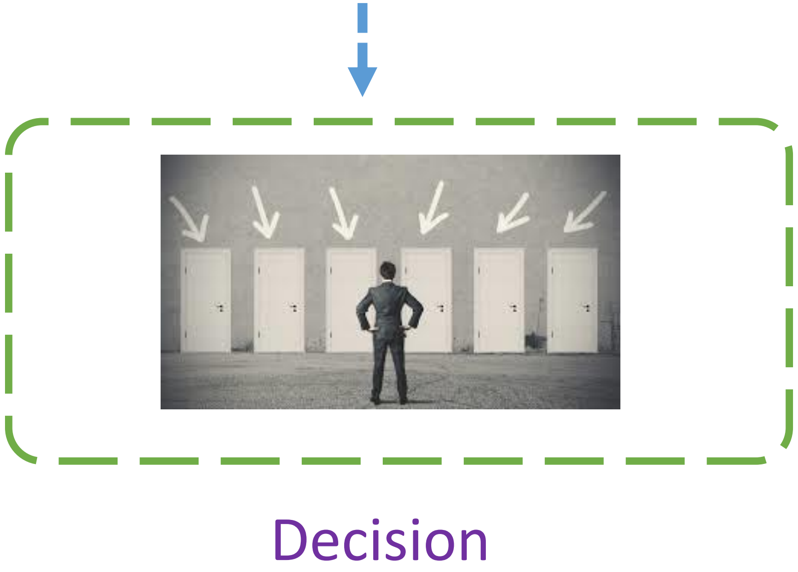
Control Database

Feed Processing Status update

Multi Stage Normalised Database

Analytics: What is Analytics ?





What is Analytics??

Analytics is the use of Math and statistics to make informed business decisions based on available information

There is an argument put forth: If the sample size gets large enough, lower accuracy of the algorithms can be acceptable

Analytics: What is Analytics ?

3 Types of Analytics

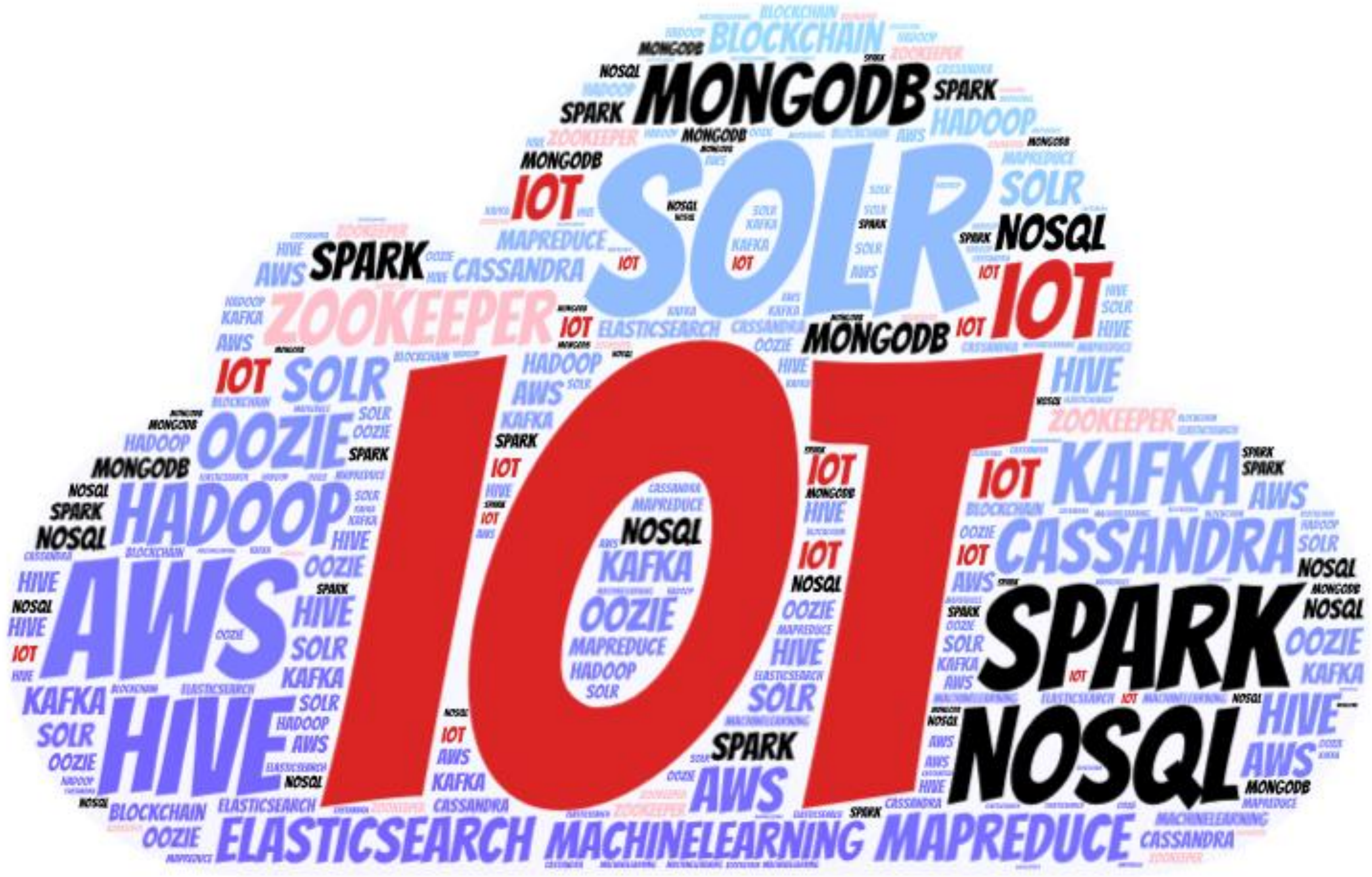
- ▶ Diagnostic\Descriptive
- ◀ Predictive
- ◀▶ Prescriptive

Applications and Benefits

- Insights on subjects we could only imagine until a few years ago.
- To uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations
- make more-informed business decisions.
- Aadhar card DBT saved thousands of crores for the Indian exchequer.
- Cash deposited under DeMo is being analyzed to find tax evaders
- Amazon understands, which of its servers will get overloaded at what time of the day.

and many more…………….

The Future : Associated Technologies



Summary: HBR Video

https://www.youtube.com/watch?v=jUFPFtU-564

Credits: HBR and Youtube

THANK YOU