# DICE ANALYTICS

***Empowering Data Analytics Ecosystem***

https://www.facebook.com/diceanalytics/

https://www.linkedin.com/company/13294896/

DICE ANALYTICS

# Introduction

**Name**
**Education**
**Organization**
**Experience**
**Expectations?**

DICE ANALYTICS

# Documentaries

https://www.youtube.com/watch?v=l6oKriR-RjM

https://www.youtube.com/watch?v=en2ix9f8ceM&list=PLBE30C2B39FE4BD1C

https://www.youtube.com/watch?v=I-SVN3txo_4

# Big Data

DICE ANALYTICS

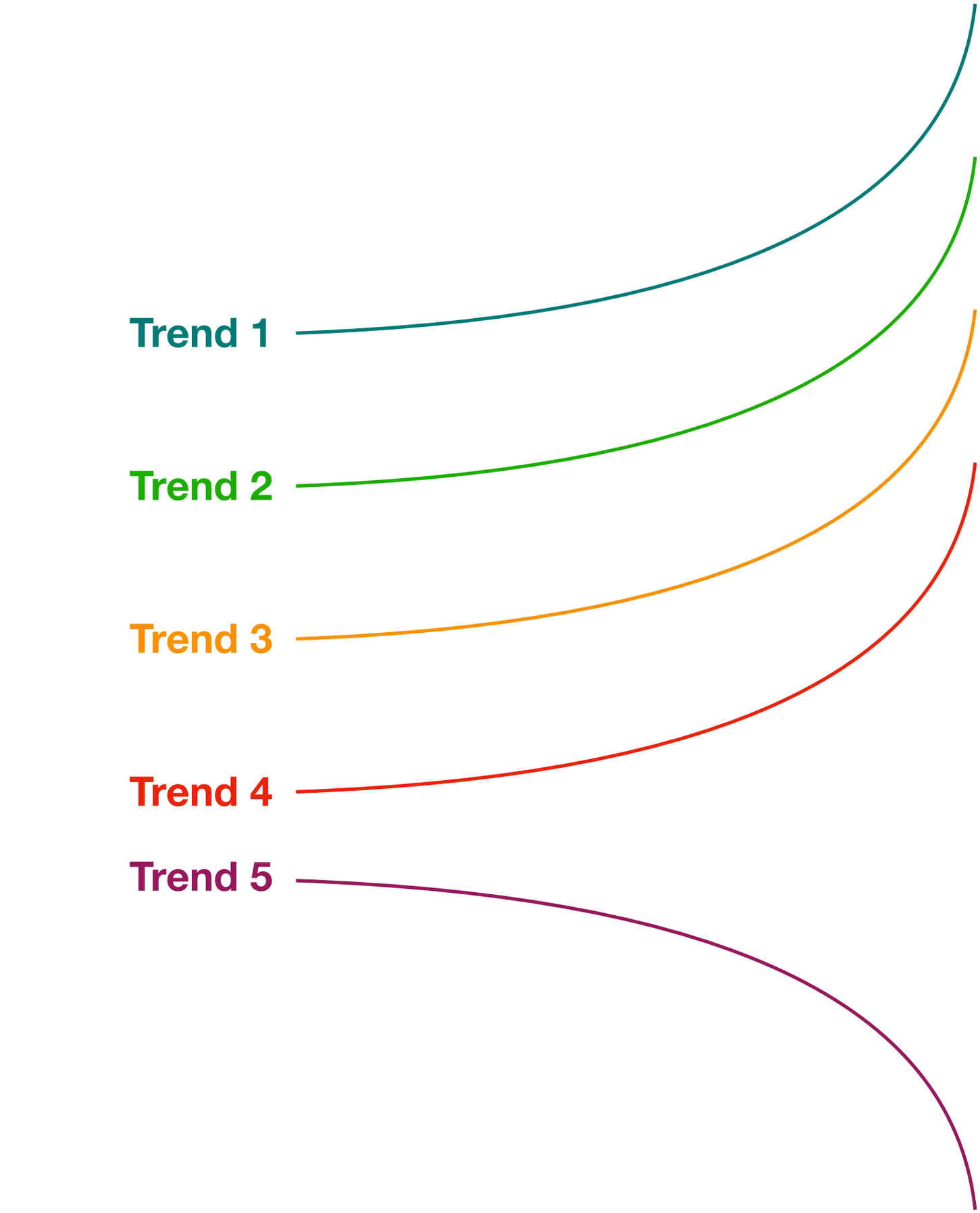
# What is Big Data?

- 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.

- Big Data challenges include capturing data, data storage, processing, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy/security.

**We will look at all these aspects in this course!**

DICE ANALYTICS




**Can you identify these trend lines in the field of data?**

DICE ANALYTICS

Trend 1

Trend 2

Trend 3

Trend 4

Trend 5

VOLUME

+

VARIETY

+

VELOCITY

+

VERACITY

+

COST

=

BIG DATA

DICE ANALYTICS

VOLUME

VARIETY

VELOCITY

File System

Databases

Data Warehouse

Big Data

VERACITY

COST

# *Big Data*

## Process A lot of Data

## High Speed

## With Less Cost

DICE ANALYTICS


# Problem Statement

DICE ANALYTICS

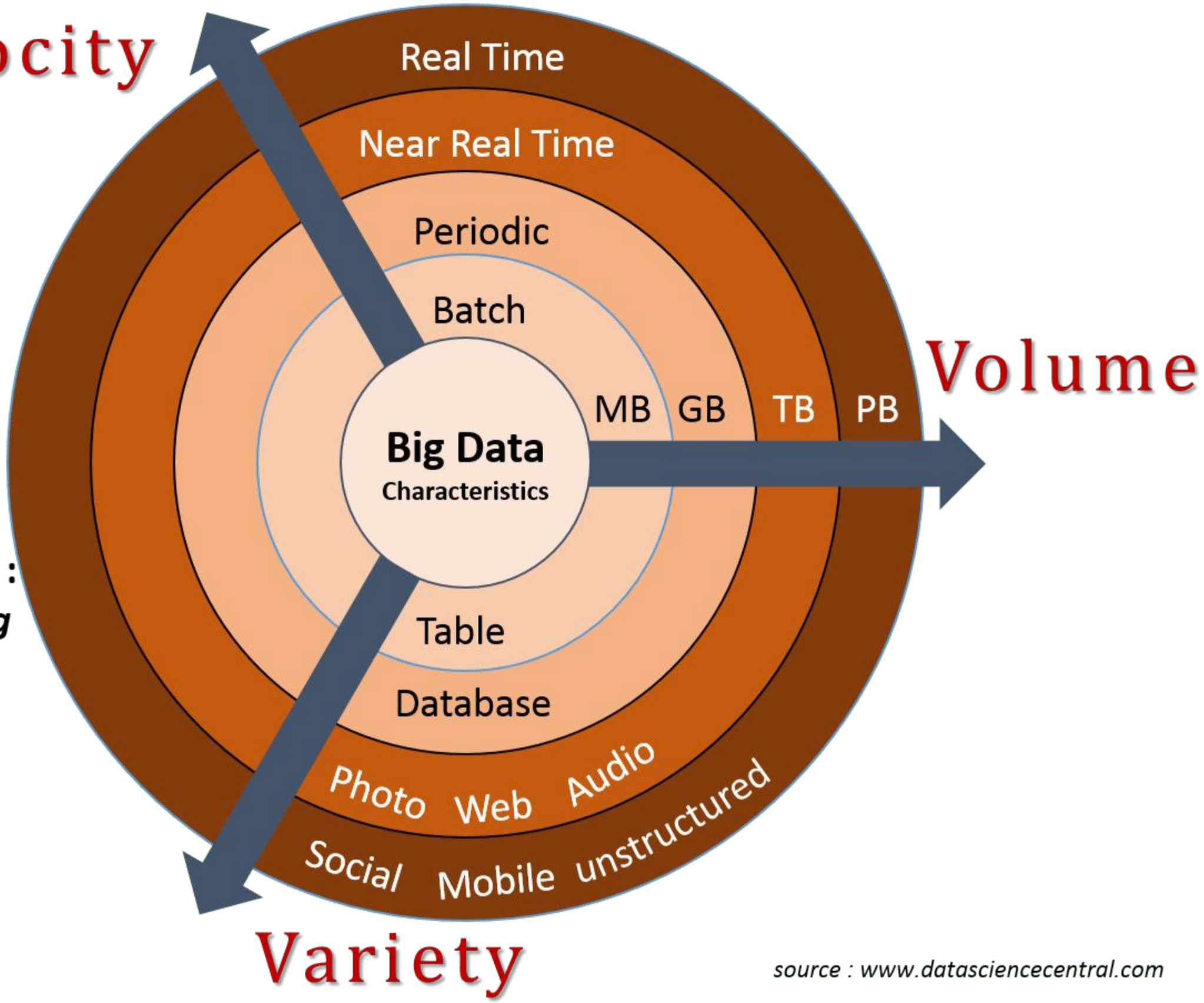
# What is Big Data?

- We often use the concept of 4 V-s to describe Big Data:

1. Volume - Amount of Data (Petabytes, Zetabytes)
2. Variety - Forms of Data (Structured, Unstructured)
3. Velocity - Speed of Data (GBs/sec)
4. Veracity - Uncertainty of Data (Accuracy)

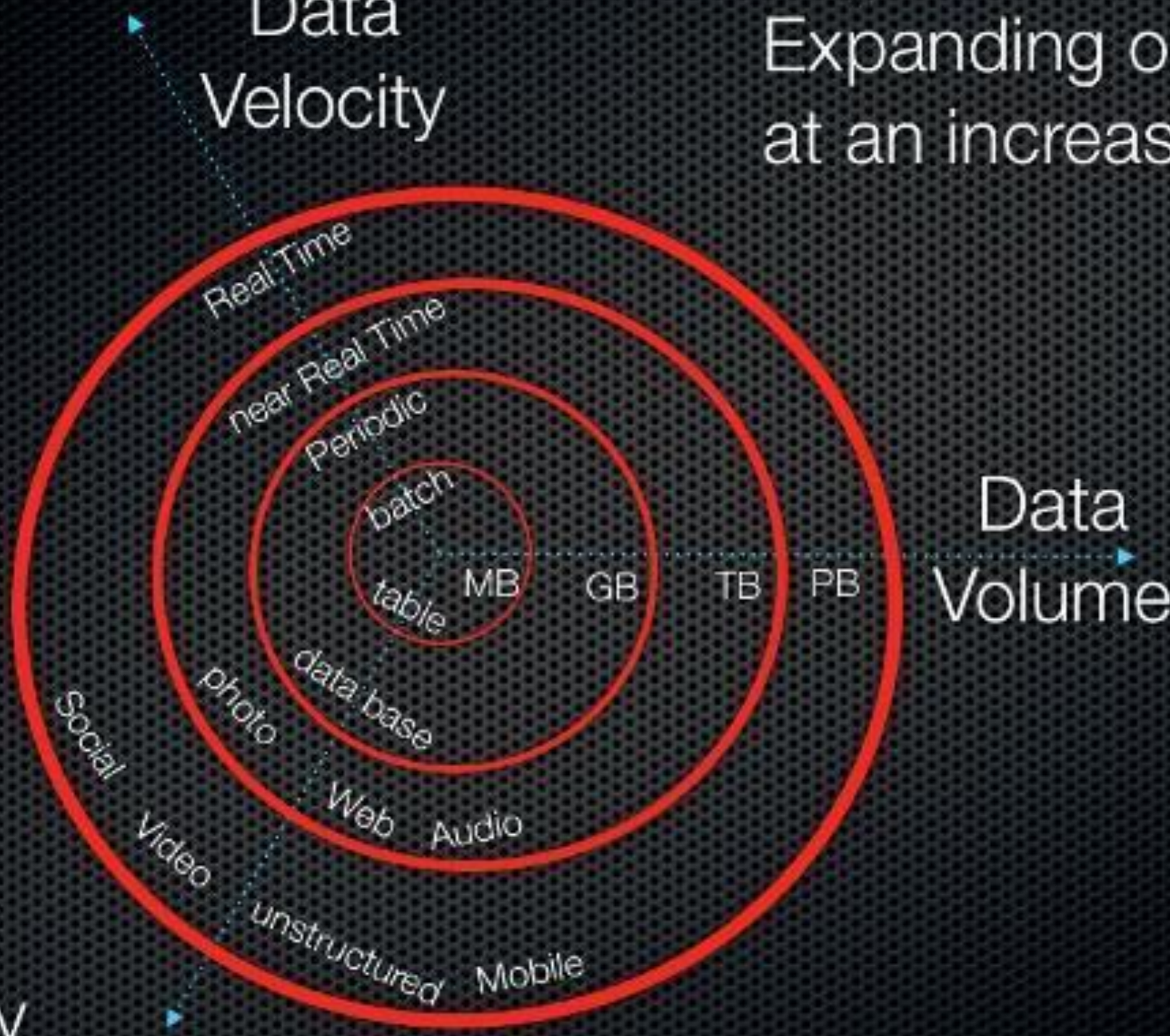
**http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg**

**Big Data Characteristics :** ***Ever Expanding horizon***

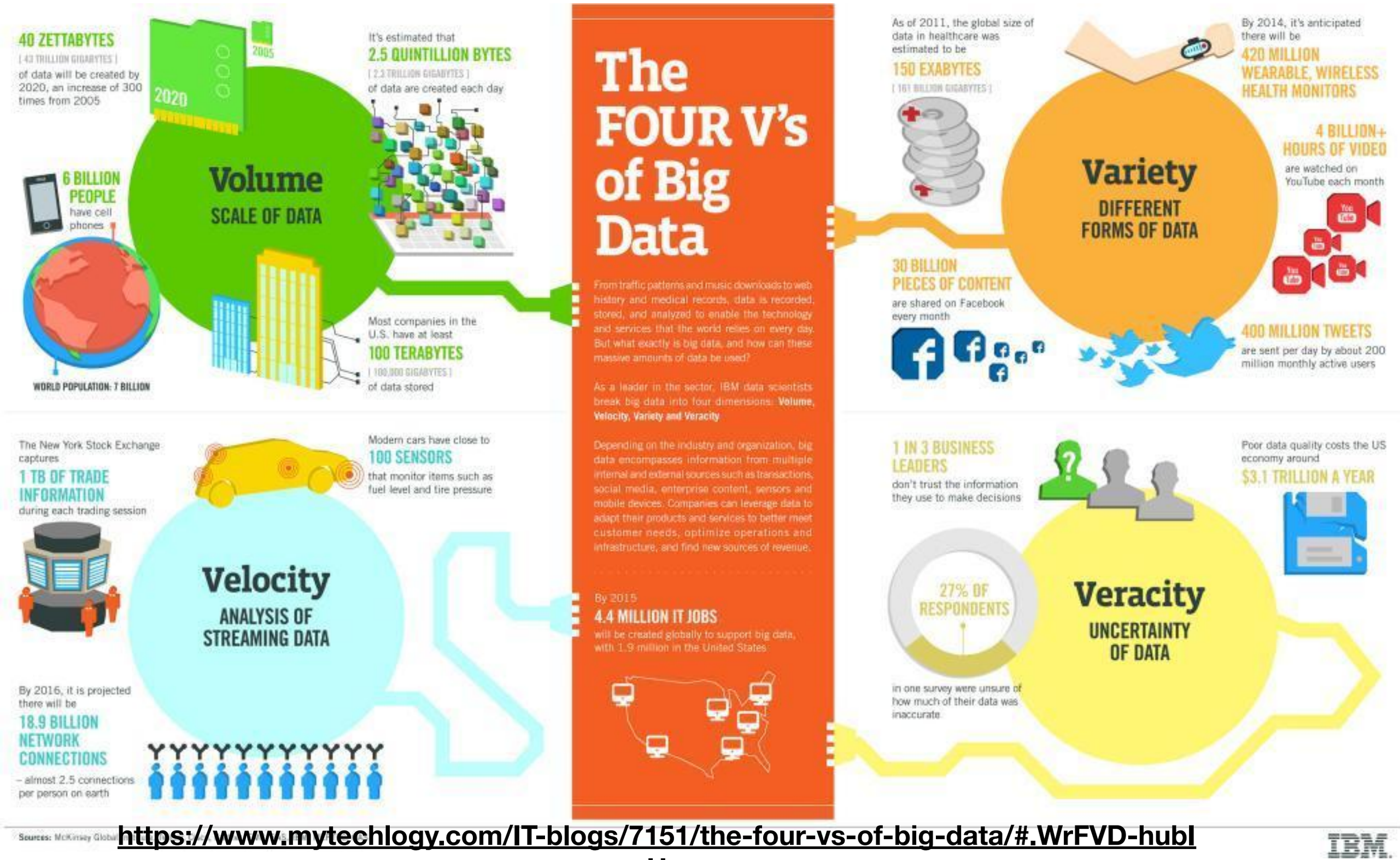



*source : www.datasciencecentral.com*

# Big Data:
# Expanding on 3 fronts at an increasing rate.



# What is Big Data?



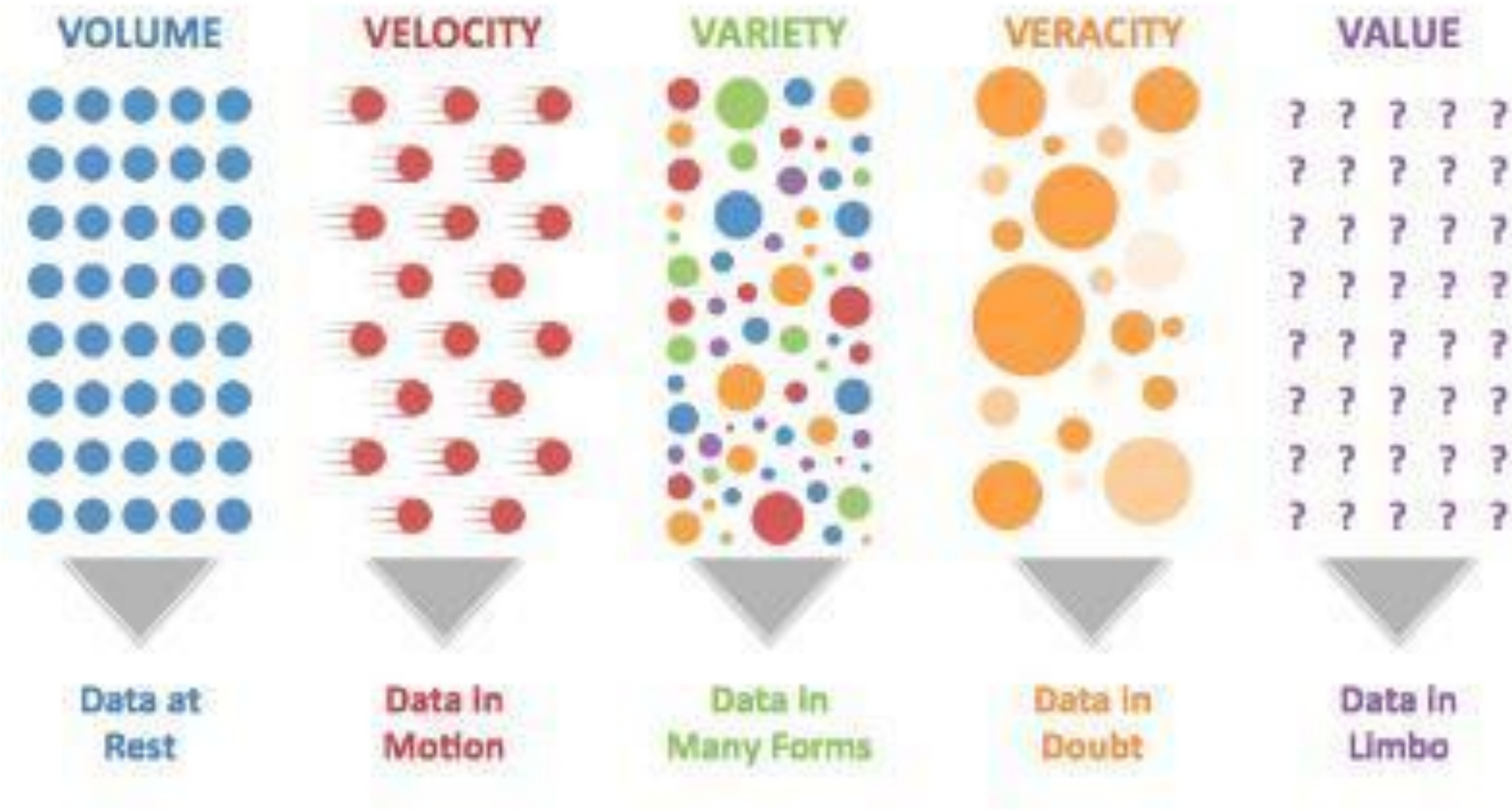

https://www.mytechlogy.com/IT-blogs/7151/the-four-vs-of-big-data/#.WrFVD-hubIU

# *What is Big Data*

**https://www.youtube.com/watch?v=tkOwlXUaGM**

**https://www.youtube.com/watch?v=xTVmK22ugj0**

**https://www.youtube.com/watch?v=Hv397JnNWYc**




http://www.evariant.com/blog/big-data-analytics-in-healthcare

# 1. Volume

- Transaction-based data stored through years.

Telcos, banks, health, etc.

Data Warehousing

- Unstructured data streaming from social media.

Images

Audio

Videos

Tweets

- Sensor and machine-to-machine data.

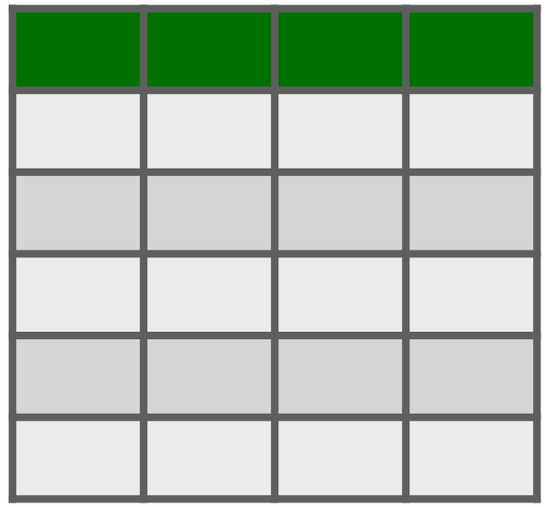
IoT

M2M

DICE ANALYTICS

# 2. Variety

- Structured data in traditional databases



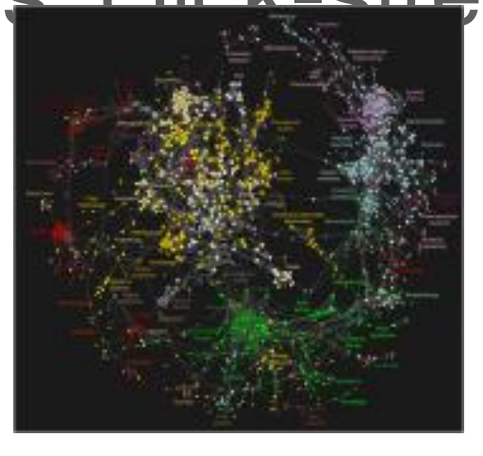
- Semi-structured data like XML or JSON.

```
"firstName": "John",
"lastName": "Smith",
"age": 25,
"address": {
    "streetAddress": "21 2nd S
    "city": "New York",
    "state": "NY",
    "postalCode": 10021
},
"phoneNumbers": [
    {
        "type": "home",
```

```
- <xml>
    <title>XML test</title>
  - <text type="test">
    - <body>
      - <p>
        Though this is a very pared
        <lb />
        down XML document, it nonetheless
        <lb />
        provides an example of how an XML
        <lb />
        document displays on the web without
        <lb />
        the intercession of a stylesheet or
        <lb />
        other conversion program.
      </p>
    </body>
  </text>
</xml>
```

- Unstructured data like emails, images, click-stream

```
STATUS | monitor  | 2012/11/11 00:51:23 | --> Monitor
STATUS | monitor  | 2012/11/11 00:51:23 | Launching a
INFO   | buserver | 2012/11/11 00:51:24 |
INFO   | buserver | 2012/11/11 00:52:09 | Nov 11, 201
INFO   | buserver | 2012/11/11 00:52:09 | INFO: Start
INFO   | buserver | 2012/11/11 00:52:09 | Nov 11, 201
INFO   | buserver | 2012/11/11 00:52:09 | INFO: Start
INFO   | buserver | 2012/11/11 00:52:09 | Nov 11, 201
INFO   | buserver | 2012/11/11 00:52:09 | WARNING: Co
2_1.xsd
INFO   | buserver | 2012/11/11 00:52:10 | Nov 11, 201
INFO   | buserver | 2012/11/11 00:52:10 | WARNING: Co
2_1.xsd
INFO   | buserver | 2012/11/11 00:52:12 | Nov 11, 201
onfig
```



# 2. Variety (cont.)

Structural Variety
Formats & Models

Semantic Variety
How to interpret and operate on data

Media Variety
Medium in which data is delivered

Availability Variety
Real-time? batch? Intermittent?

# 3. Velocity

- Megabytes per second, Gigabytes per second.


- Data needs to be dealt with in timely manner.



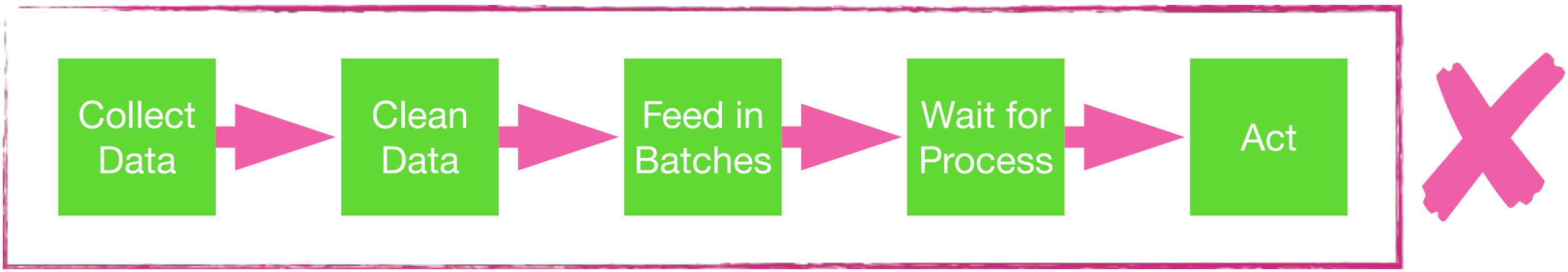
- Inconsistent data flows with periodic peaks.

**Speed of Creating Data**
**Speed of Storing Data**
**Speed of Processing Data**
**Speed of Analyzing Data**


DICE ANALYTICS

# 3. Velocity (cont.)

**Batch Processing**




**Real-Time Processing**

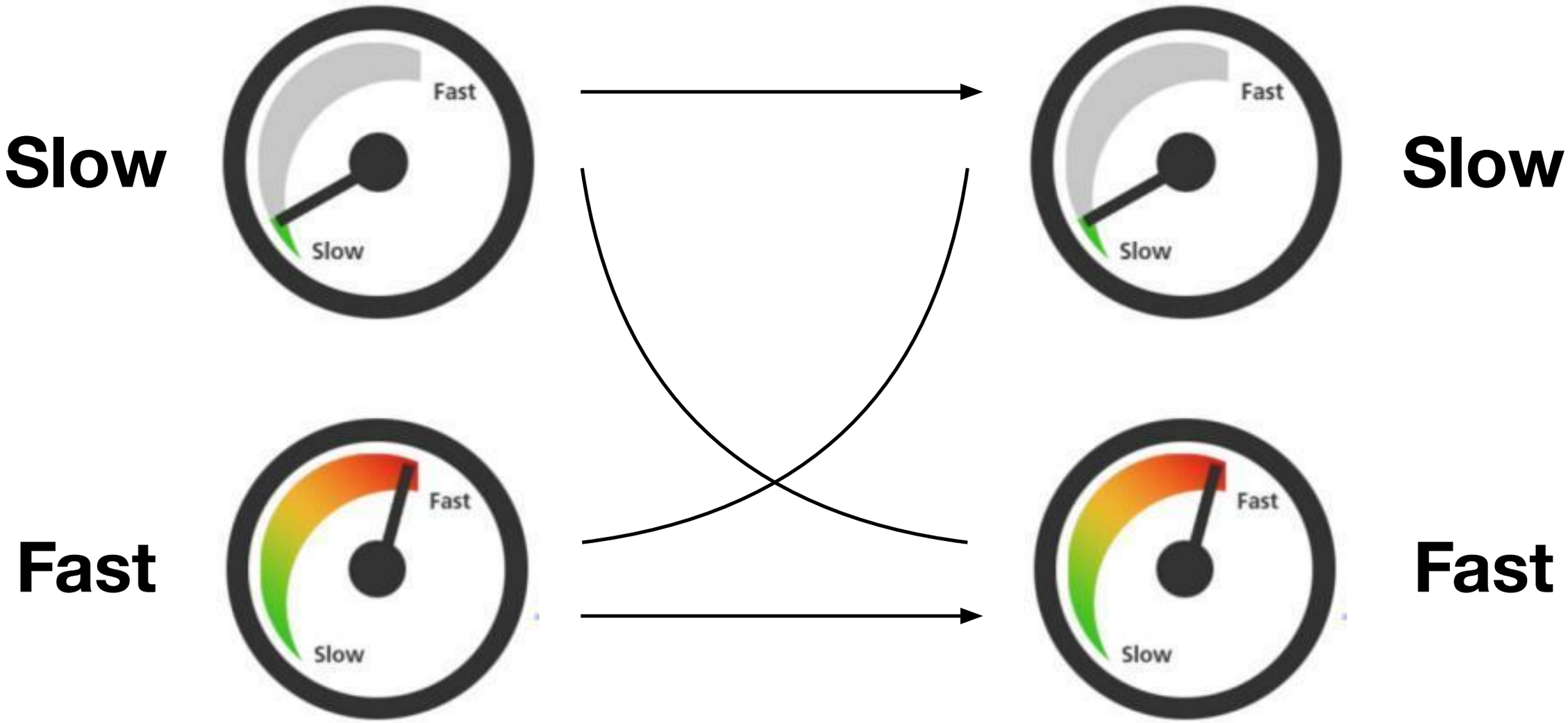

**Big Data enables real-time decision pipelines!**

DICE ANALYTICS

# 3. Velocity (cont.)

**Speed of Data Generation**

**Speed of Data Processing**

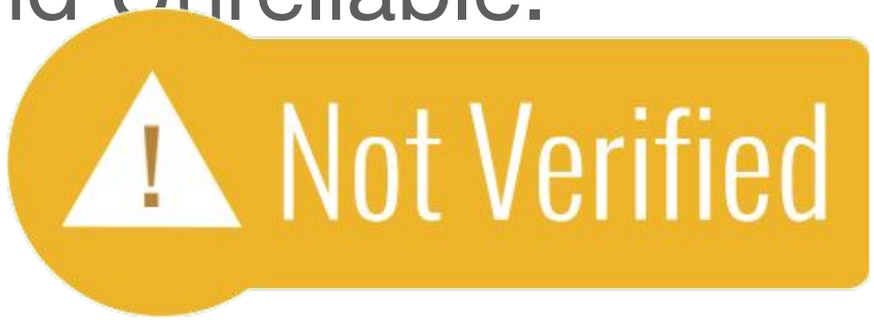



**Which path to choose in what scenario?**

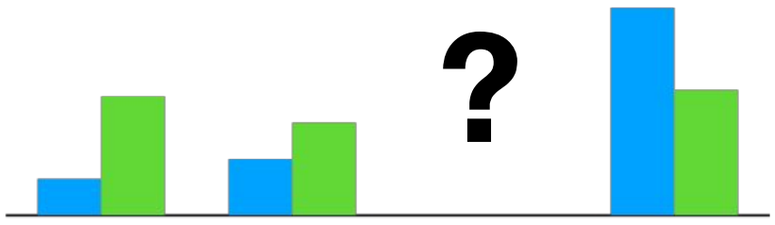
DICE ANALYTICS

# 4. Veracity

- Untrusted and Unreliable.



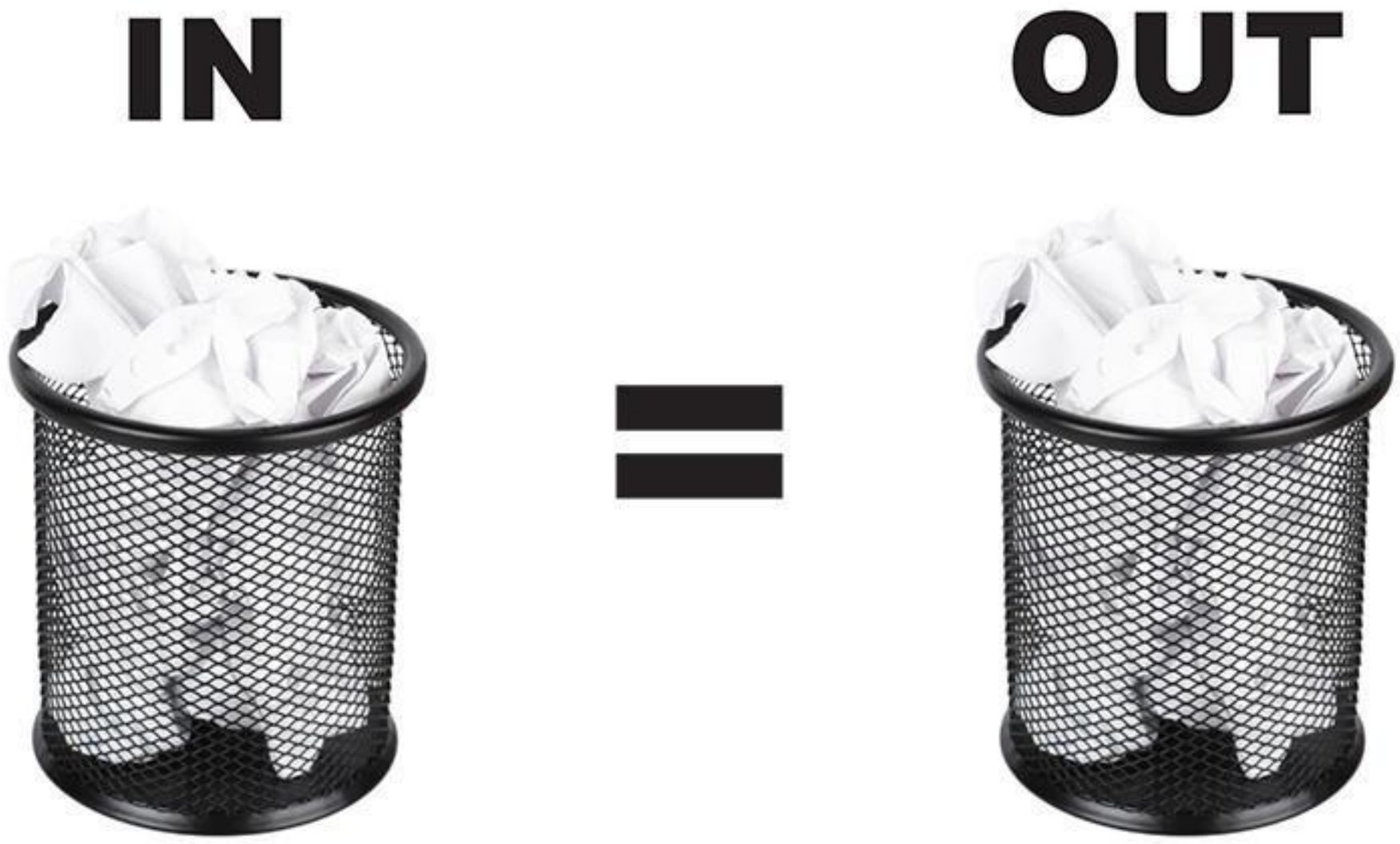

- Data Inconsistency and Incompleteness.




- Biased, Unclean and Ambiguous Data


DICE ANALYTICS

# 4. Veracity (cont.)



# Which of the Following Organization is Facing Big Data Problem ?

FBR PAKISTAN

Jazz

Rahat Since 1950

DICE ANALYTICS

# *Does Every Organization Faces Big Data Problem*

# *NO*

DICE ANALYTICS

# *Big Data Problem*
# *&*
# *Big Data Platforms*

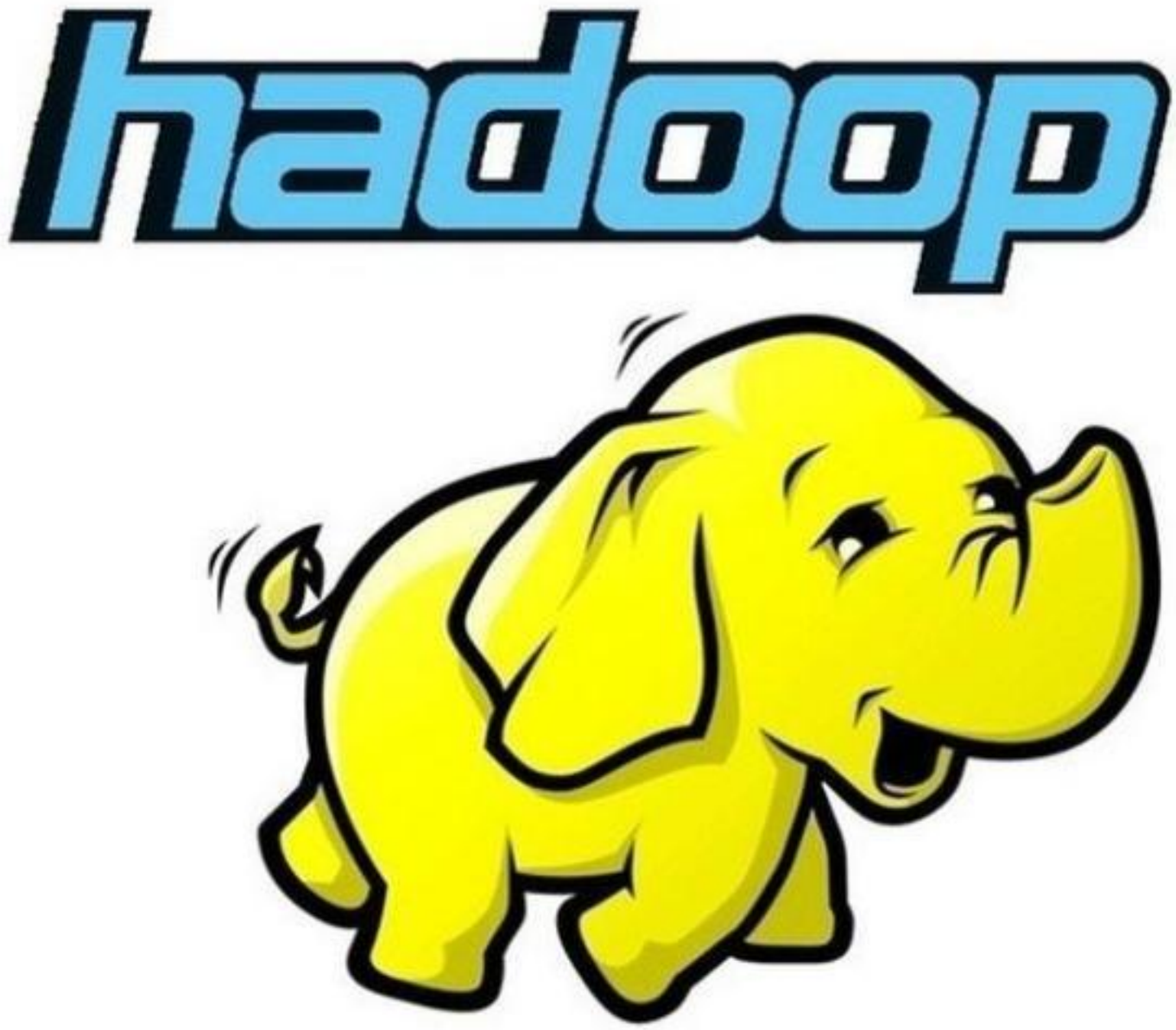
DICE ANALYTICS

# We Buy Machines

## Storage

## Processing

DICE ANALYTICS

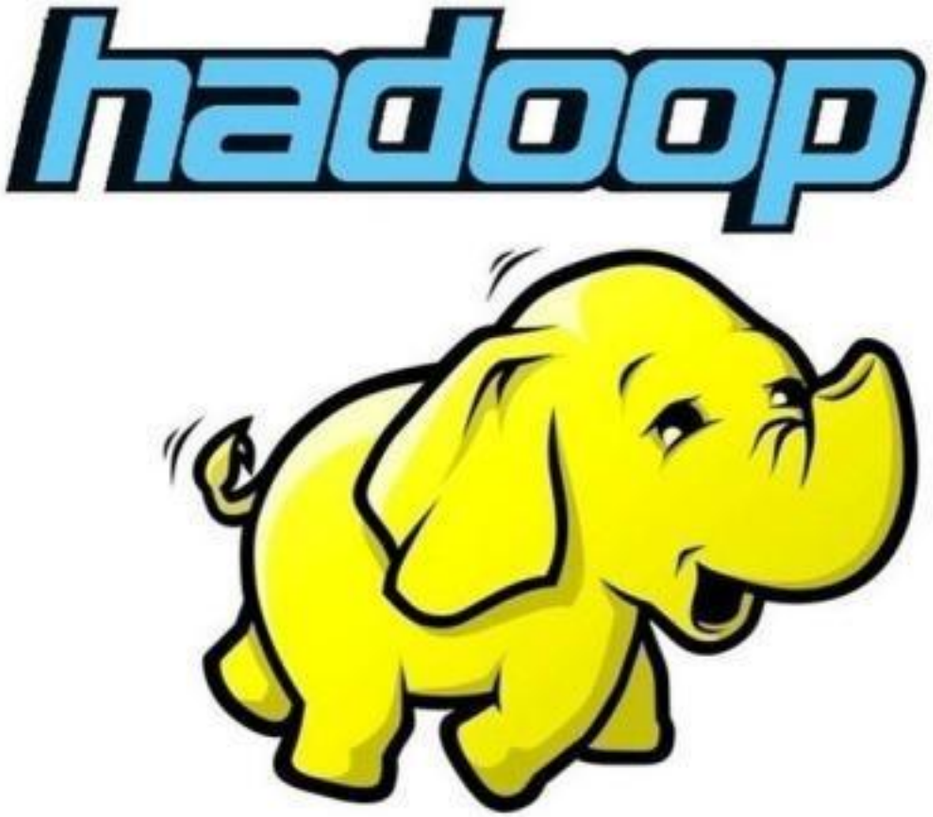


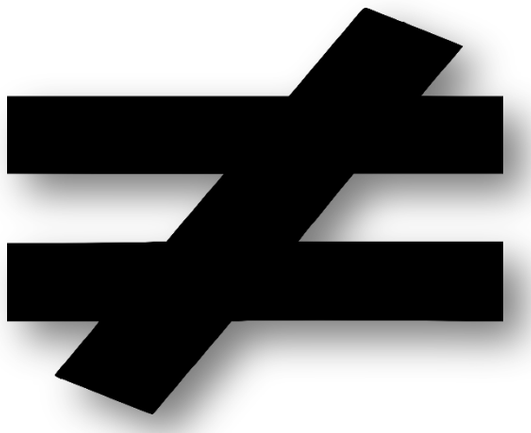

# A Big Data Platform

DICE ANALYTICS


# A Big Data Platform

DICE ANALYTICS

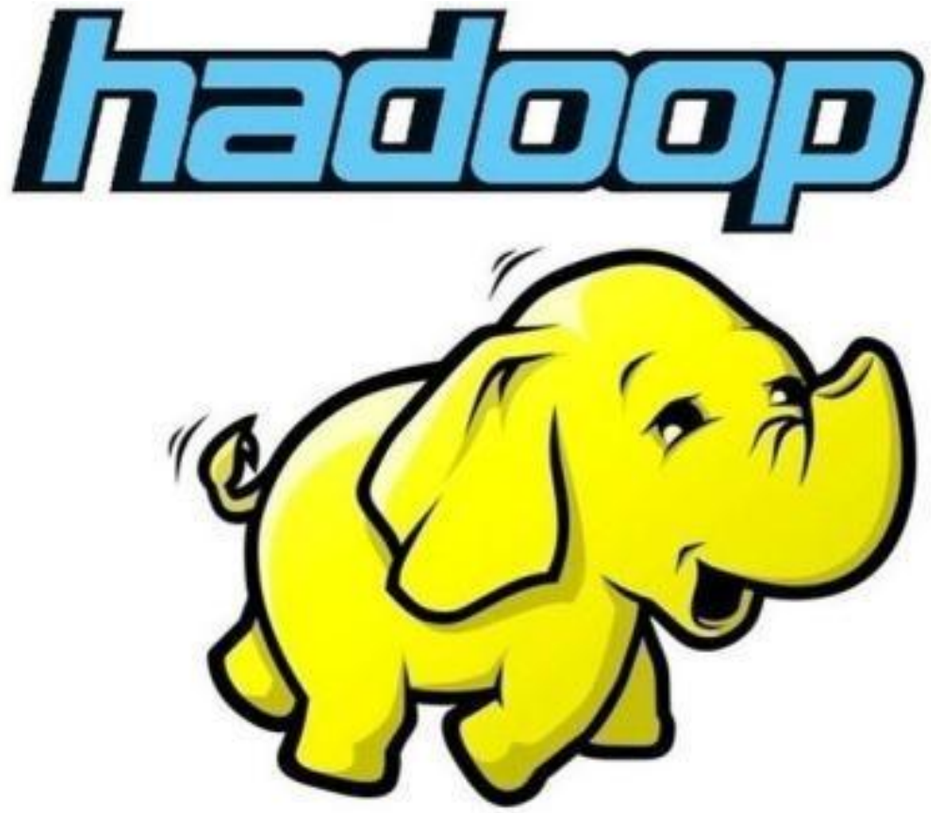



# Hadoop is one of the platform to

# Solve Big Data Problem

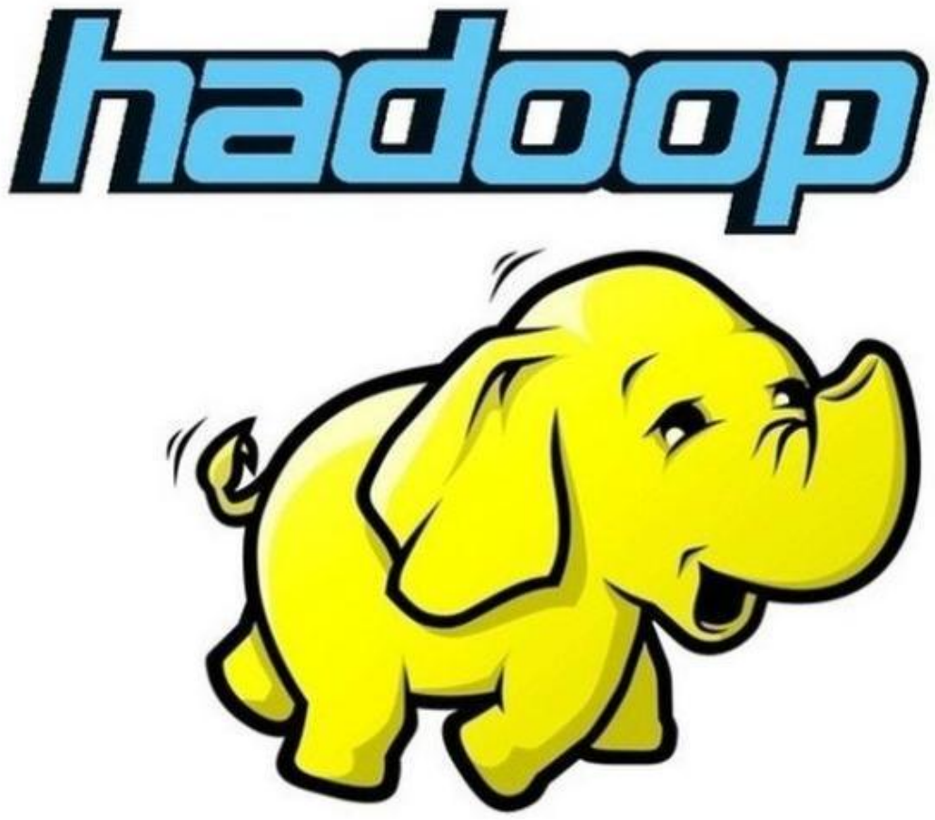
DICE ANALYTICS




# Distributed Storage

# Parallel Processing

DICE ANALYTICS





DICE
ANALYTICS

cloudera

MAPR™ TECHNOLOGIES

Hortonworks

CLOUDERA

DICE ANALYTICS




DICE ANALYTICS

# Why Big Data Platforms?

Scalable

Cost Effective

Flexible

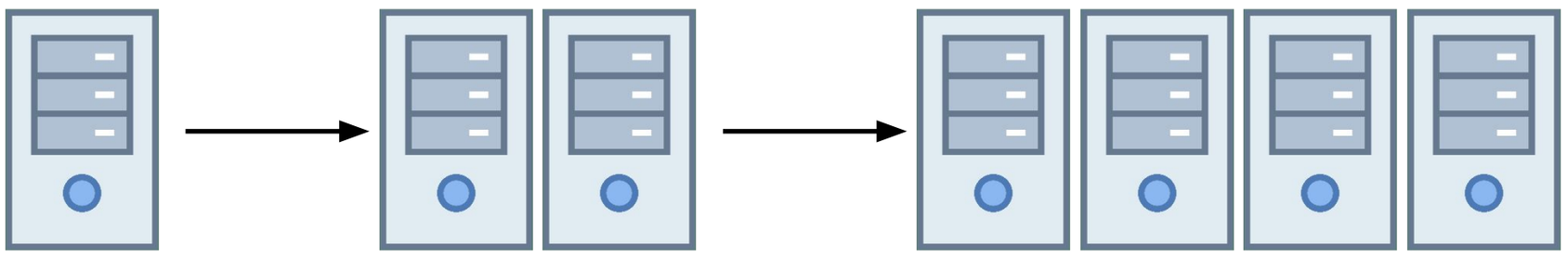
Fast

Resilient

DICE ANALYTICS

# 1. Scalable

- It can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel.
- It enables businesses to run applications on thousands of nodes involving thousands of terabytes of data.
- It manages horizontal scalability seamlessly.



DICE ANALYTICS

# 2. Cost Effective

- A scale-out architecture (as seen in previous slide) that can *affordably* store all of a company's data for later use.
- In the past, many companies would have had to down-sample data, in an effort to reduce costs.
- The raw data would be deleted in relational DBs, as it would be too cost-prohibitive to keep.


**The cost savings are staggering!**

DICE ANALYTICS

# 3. Flexible

- Enables businesses to easily access new data sources and tap into different types of data (structured, unstructured, semistructured).
- A single system deriving valuable business insights from data sources as variable as social media, email conversations or clickstream data.
- A single system used for a wide variety of purposes, such as log processing, recommendation systems, data warehousing, market campaign analysis and fraud detection.

DICE ANALYTICS

# 4. Fast

- Storage method is based on a distributed file system that basically 'maps' data wherever it is located on a cluster.

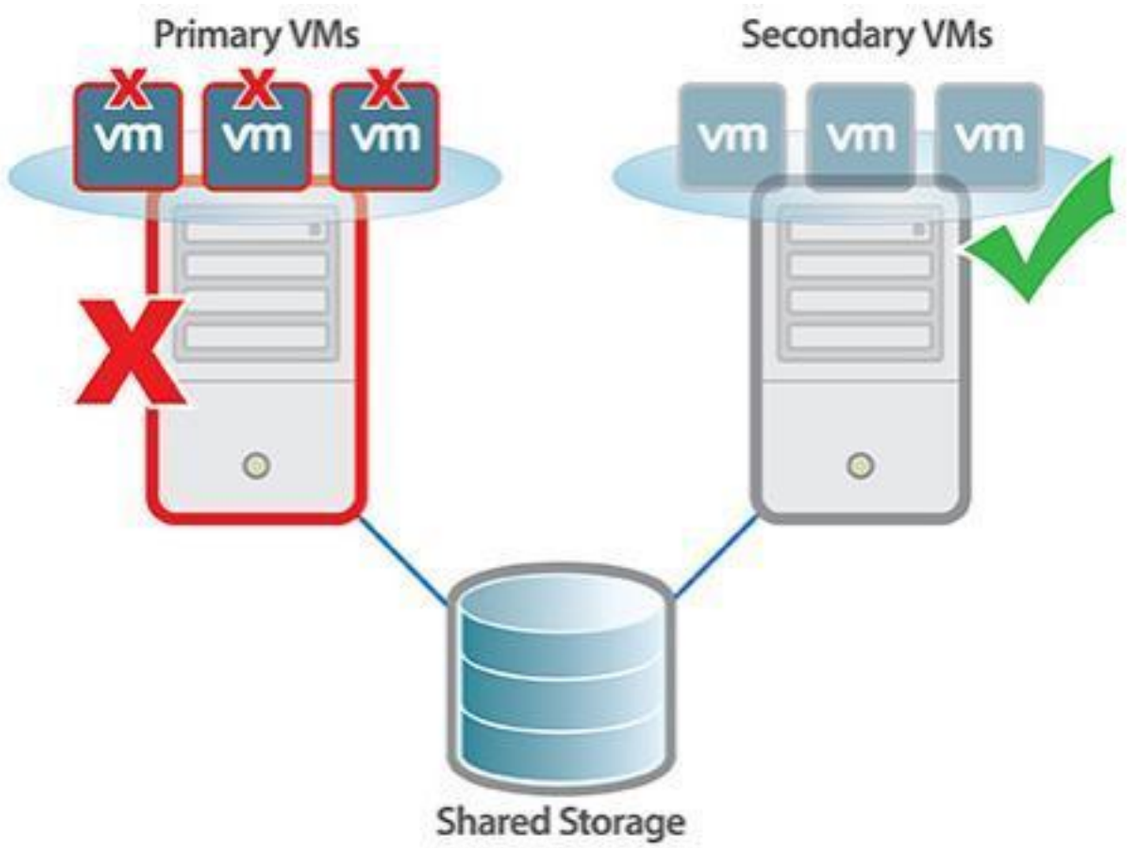
- The tools for data processing are often on the same servers where the data is located, resulting in much faster data processing.

- If you're dealing with large volumes of unstructured data, it is able to efficiently process terabytes of data in just minutes, and petabytes in hours.

DICE ANALYTICS

# 5. Resilient

- Data is replicated to many nodes in the cluster, which means that in the event of failure, there is another copy available for use.




DICE ANALYTICS

# Sources of Big Data

Machines

People

Organizations

DICE ANALYTICS

# 1. Machines

- Machine generated data is the biggest source of Big Data.

**A Boeing 787 produces 1/2 Terabytes per flight!**


- Internet of Things, Smart Devices - phones & sensors.

- Enable real-time decisions, like Fraud Detection.

‘A lot of smart devices’ x ‘A lot of data capture’ = Big Data

# 2. People

- Mostly unstructured and text-heavy.


- 80-90% of data the total data in the world is unstructured.


- 75% of total data on internet is images/videos. It’s called the Dark Matter of web.

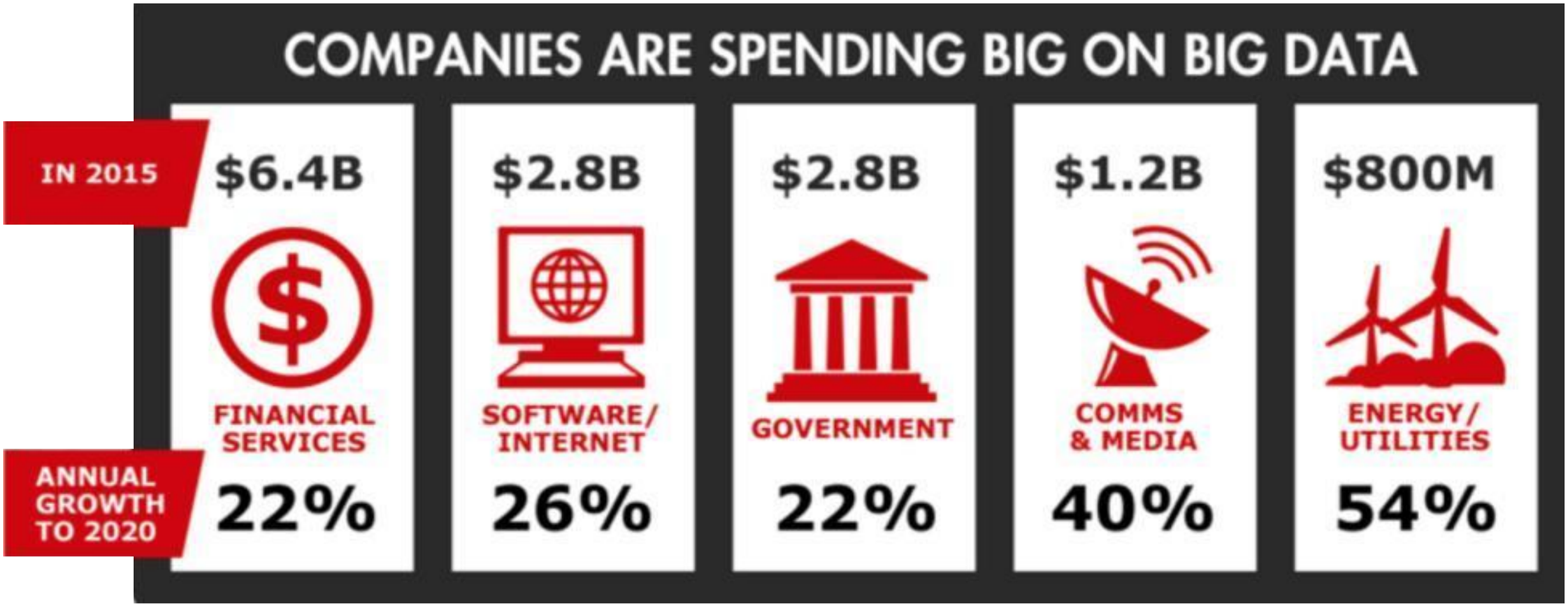
DICE ANALYTICS

# 3. Organizations

- Most data is Structured Commercial Transactions, Govt. Open Data, Banking Stock Records, Medical Health Records, E-Commerce, etc.
  - At least as important as unstructured data.
- It often gets ‘compartmentalised’ into isolated information islands called ***Data Silos***.
- Benefits can generated only by linking with other structured and non-structured data.
- Walmart collects 2.5 petabytes of data per hour!

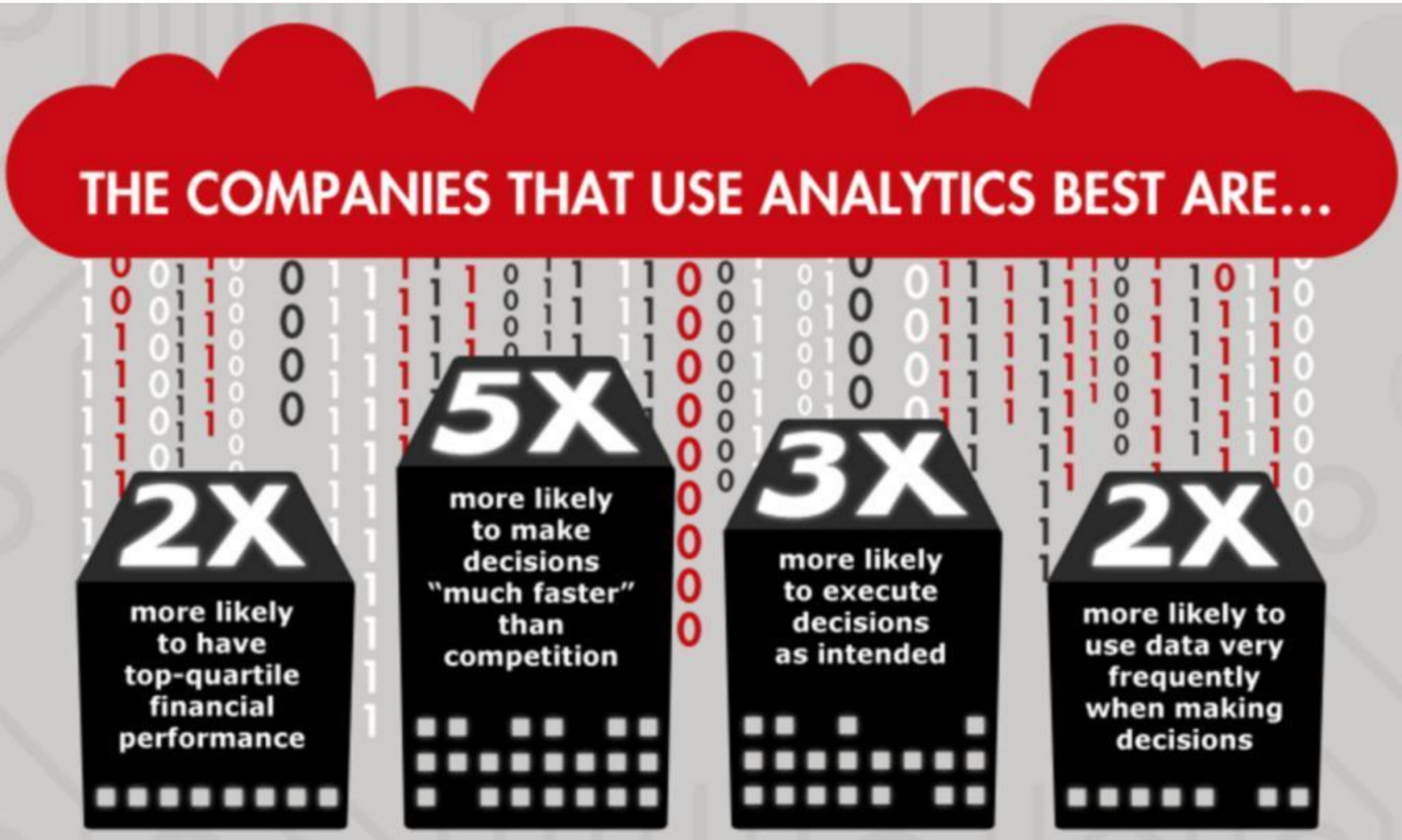

Walmart

# Intelligent Companies



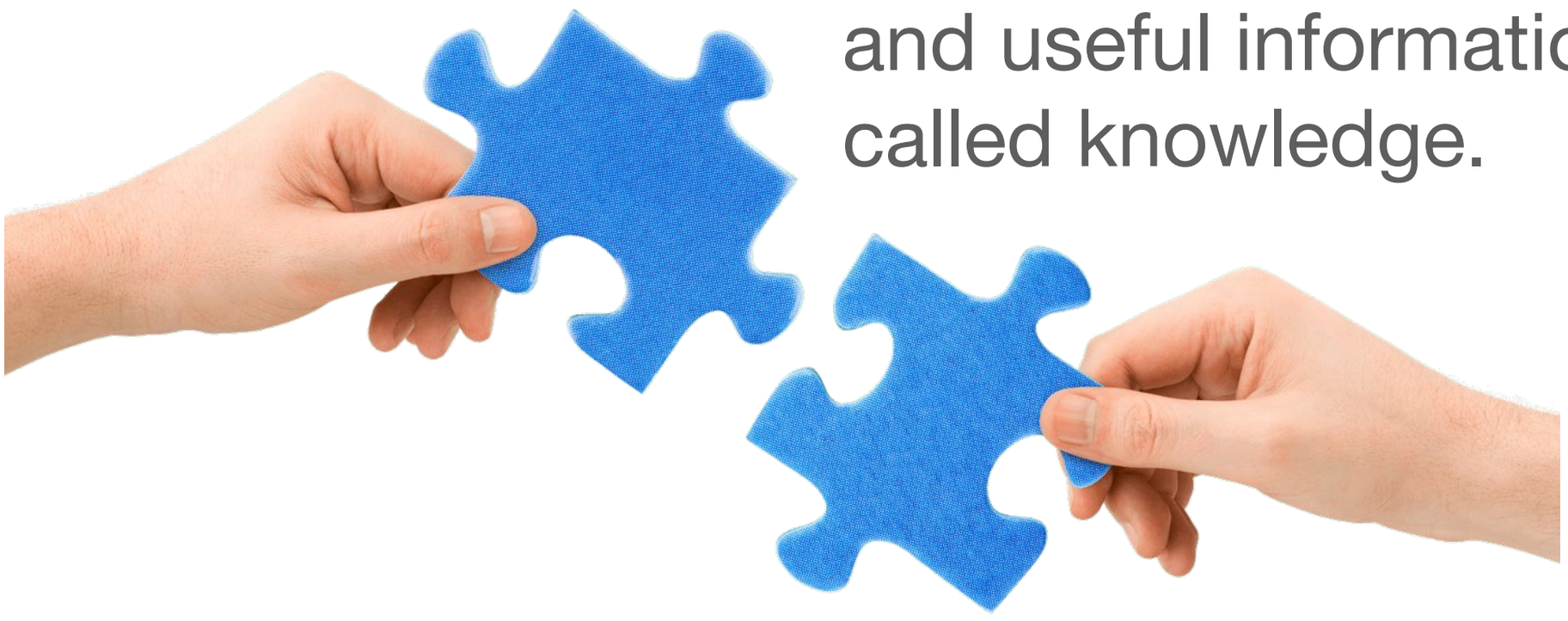

DICE ANALYTICS

# Growth


THE COMPANIES THAT USE ANALYTICS BEST ARE...
2X
more likely to have top-quartile financial performance
5X
more likely to make decisions "much faster" than competition
3X
more likely to execute decisions as intended
2X
more likely to use data very frequently when making decisions

# The Key: Integration

- The key to success is integration of diverse data!



- Bringing together data from diverse sources and turning them into coherent and useful information, called knowledge.

- Reduced data complexity & Increased data availability.

# Applications of Big Data

- Personalized Marketing.

- Recommendation Engines.

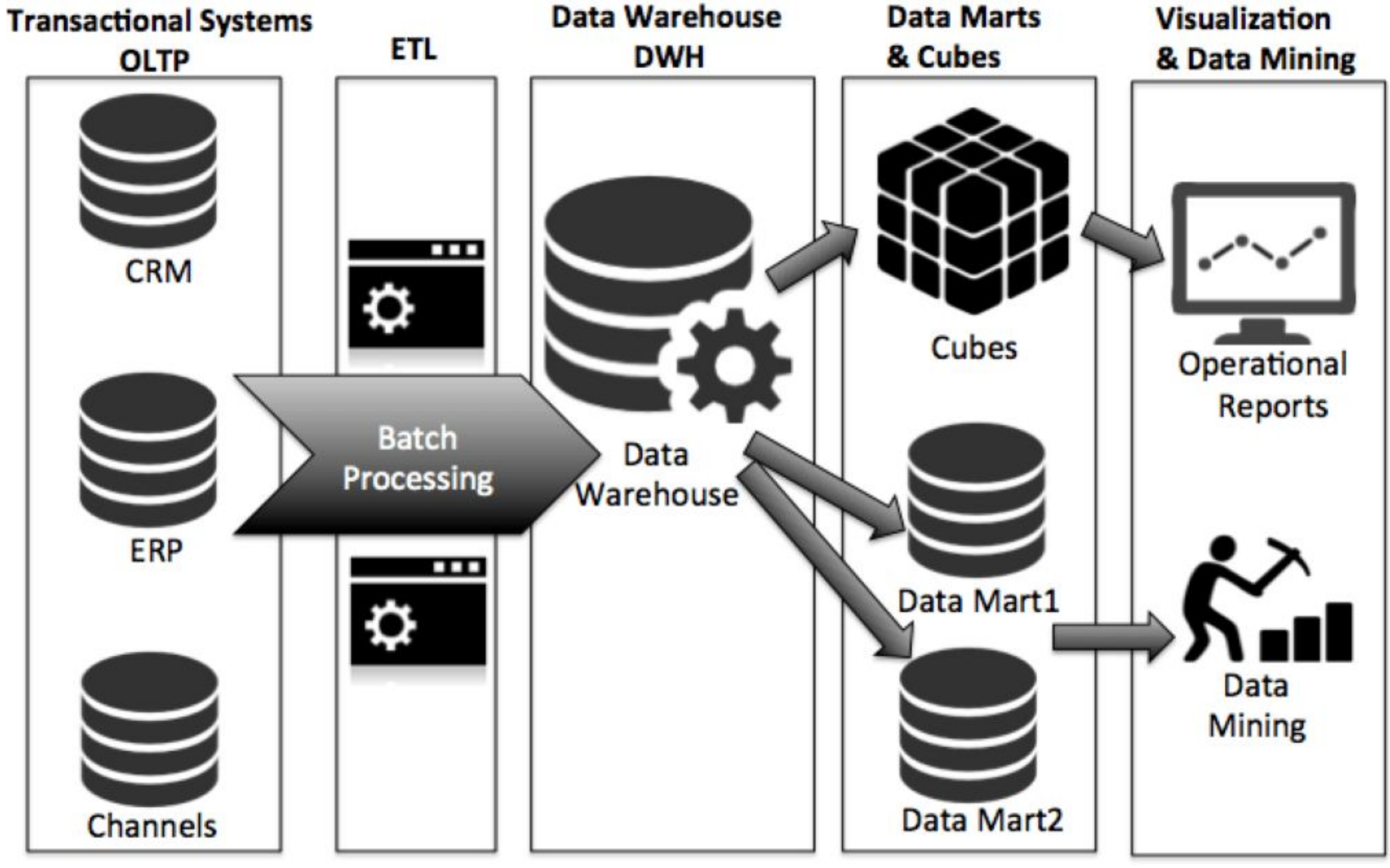
- Sentiment Analysis.

- Mobile Advertising.

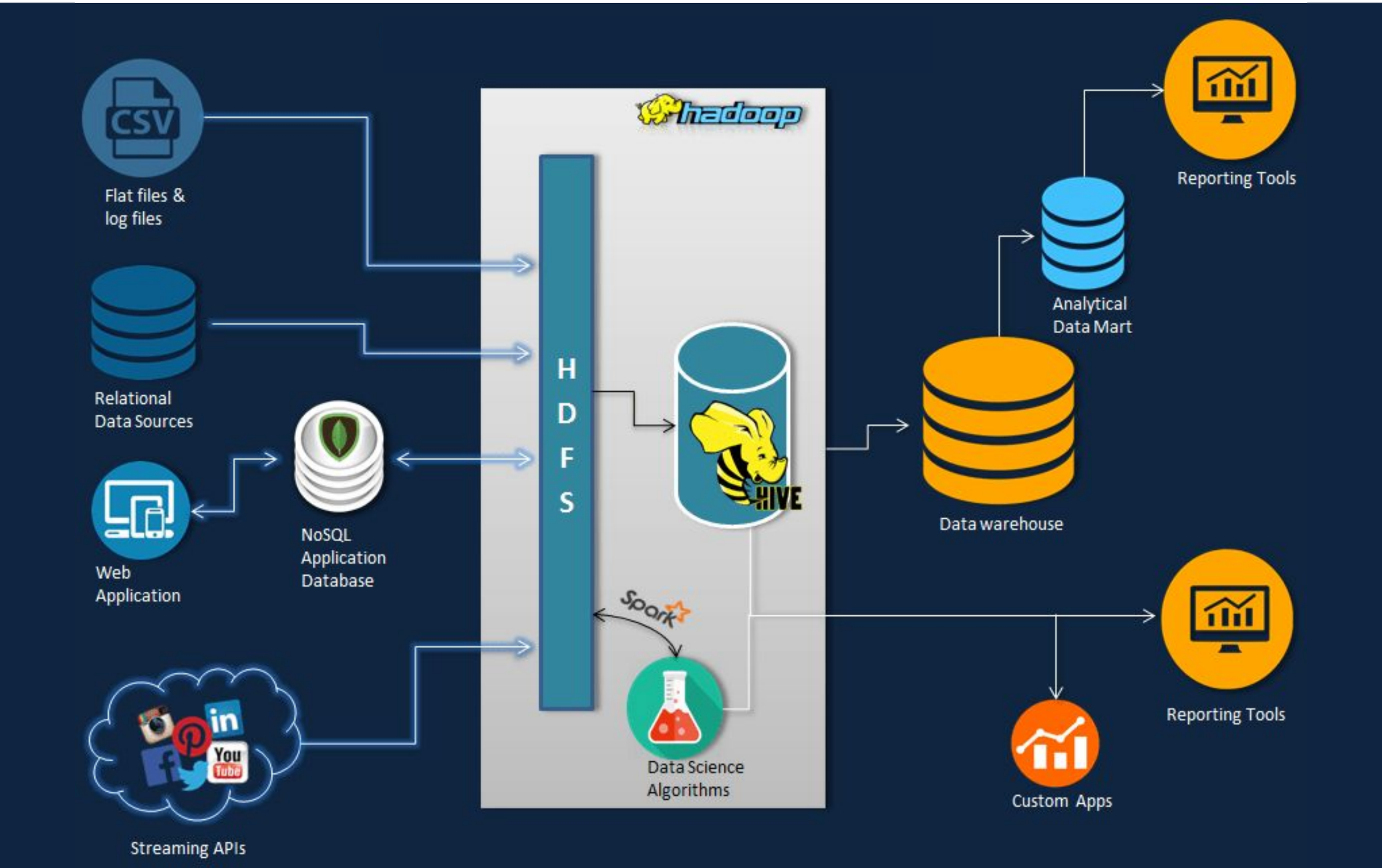
- Biomedical Applications.

- Smart Cities.

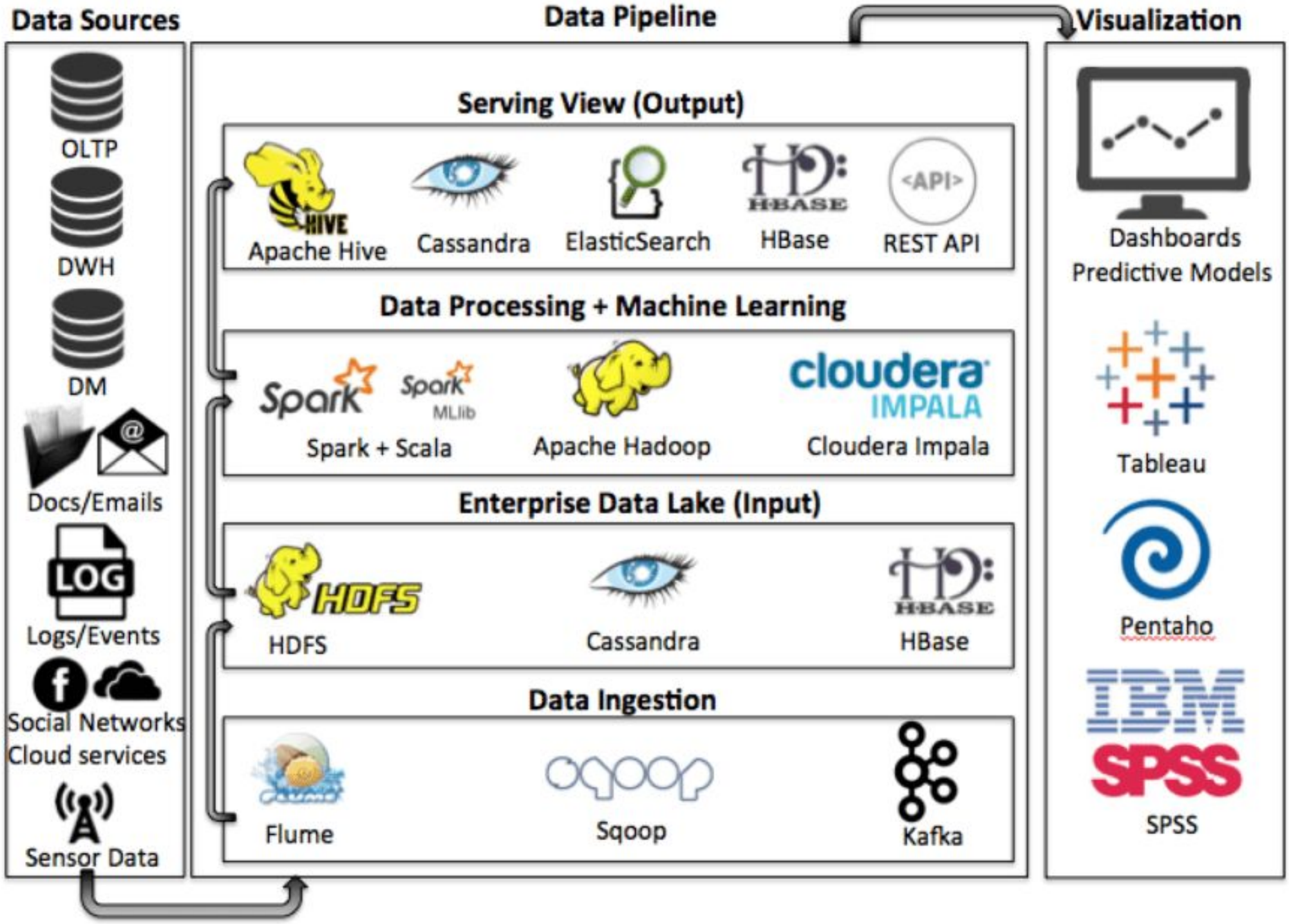
# Traditional Data Warehouse



# Modern Data Warehouse



# Modern Data Pipelines



# How to Get Value Out of Big Data?

## Data Science

DICE ANALYTICS

# Data Science vs Big Data

1)

Questio
n

Acquir
e

Ingest/ETL

Exploratory Data Analysis
(EDA)

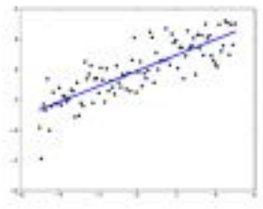
Wrangling

Visualiz
e

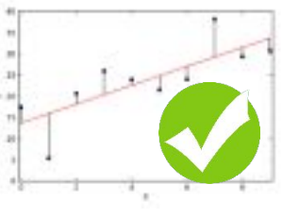
Modellin
g

2)

Choos
e



Build/Train



Validate

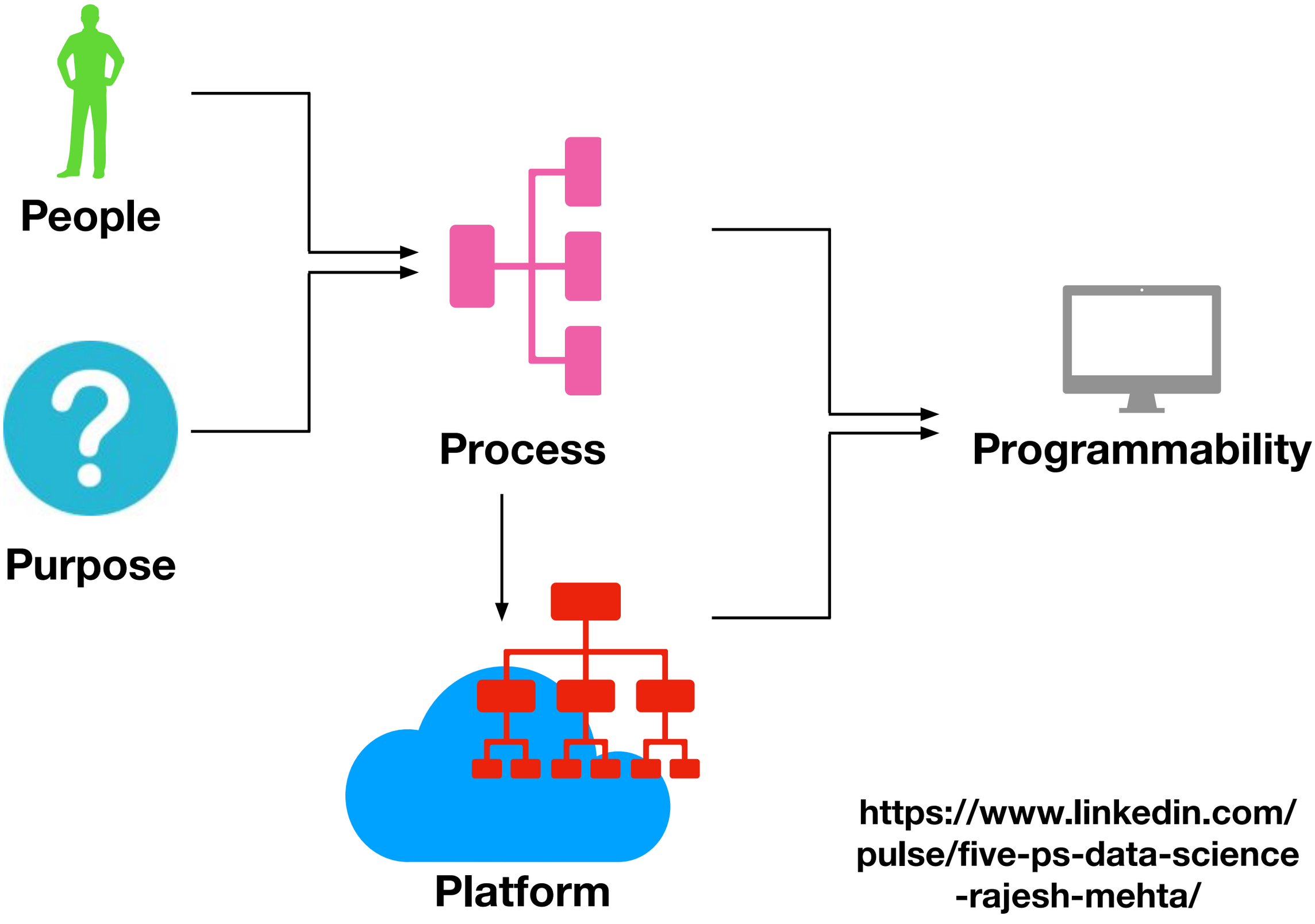
Deplo
y

Test

3)


DICE
ANALYTICS

# 5 P's of Data Science




https://www.linkedin.com/pulse/five-ps-data-science-rajesh-mehta/


DICE ANALYTICS

# What is Apache Hadoop?

- Apache Hadoop software library is a **framework** that allows for the **distributed processing** of large data sets **across clusters** of computers using simple programming models.

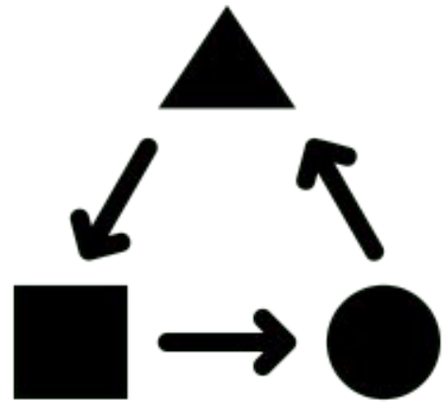
- It is designed to **scale up** from single servers to thousands of machines, each offering **local computation and storage**.

- Rather than rely on hardware to deliver **high-availability**, the library itself is designed to detect and **handle failures** at the application layer.

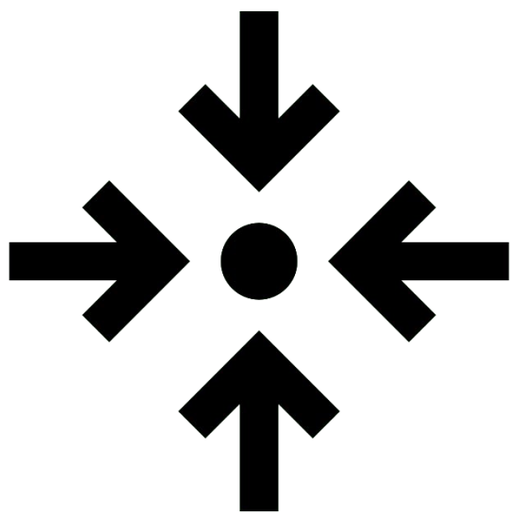
DICE ANALYTICS

# Open Enterprise Hadoop

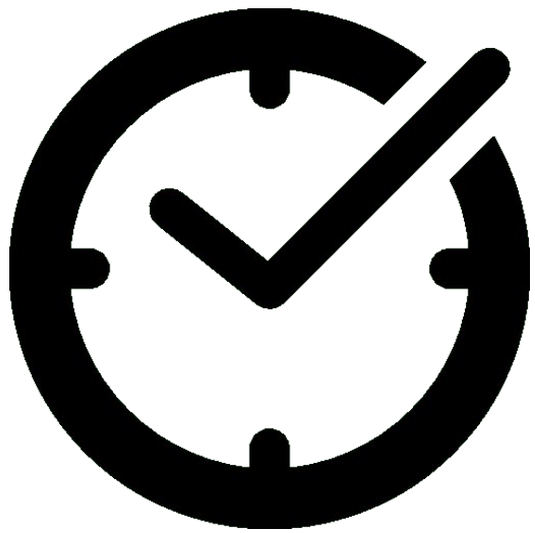
**Open**



**Interoperable**

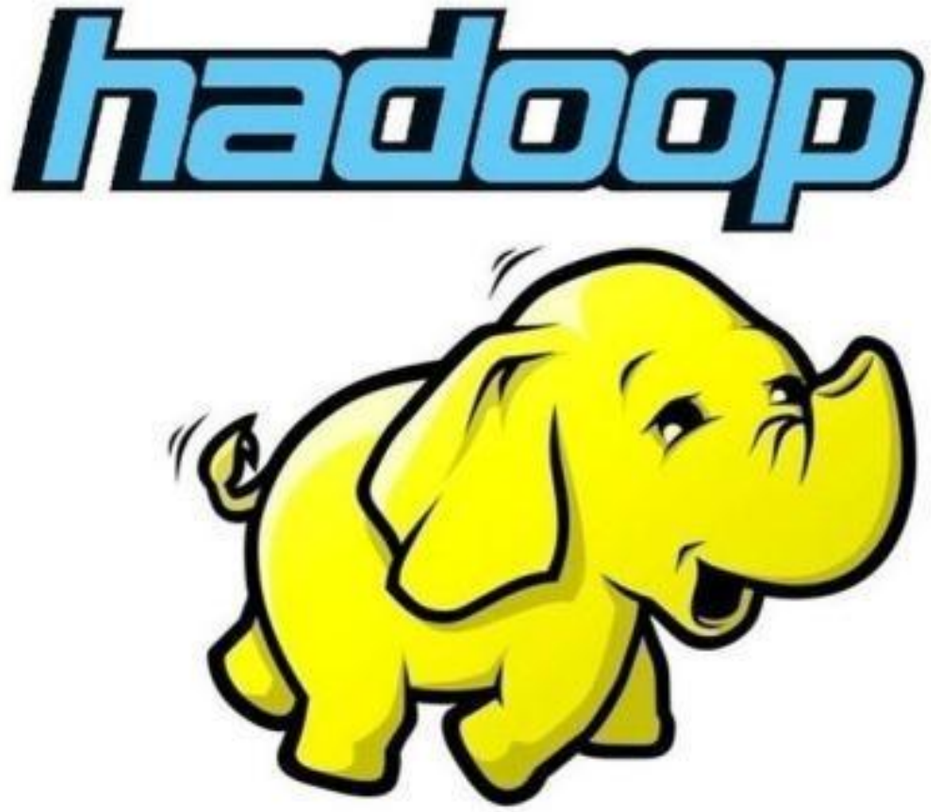


**Central**



**Available**

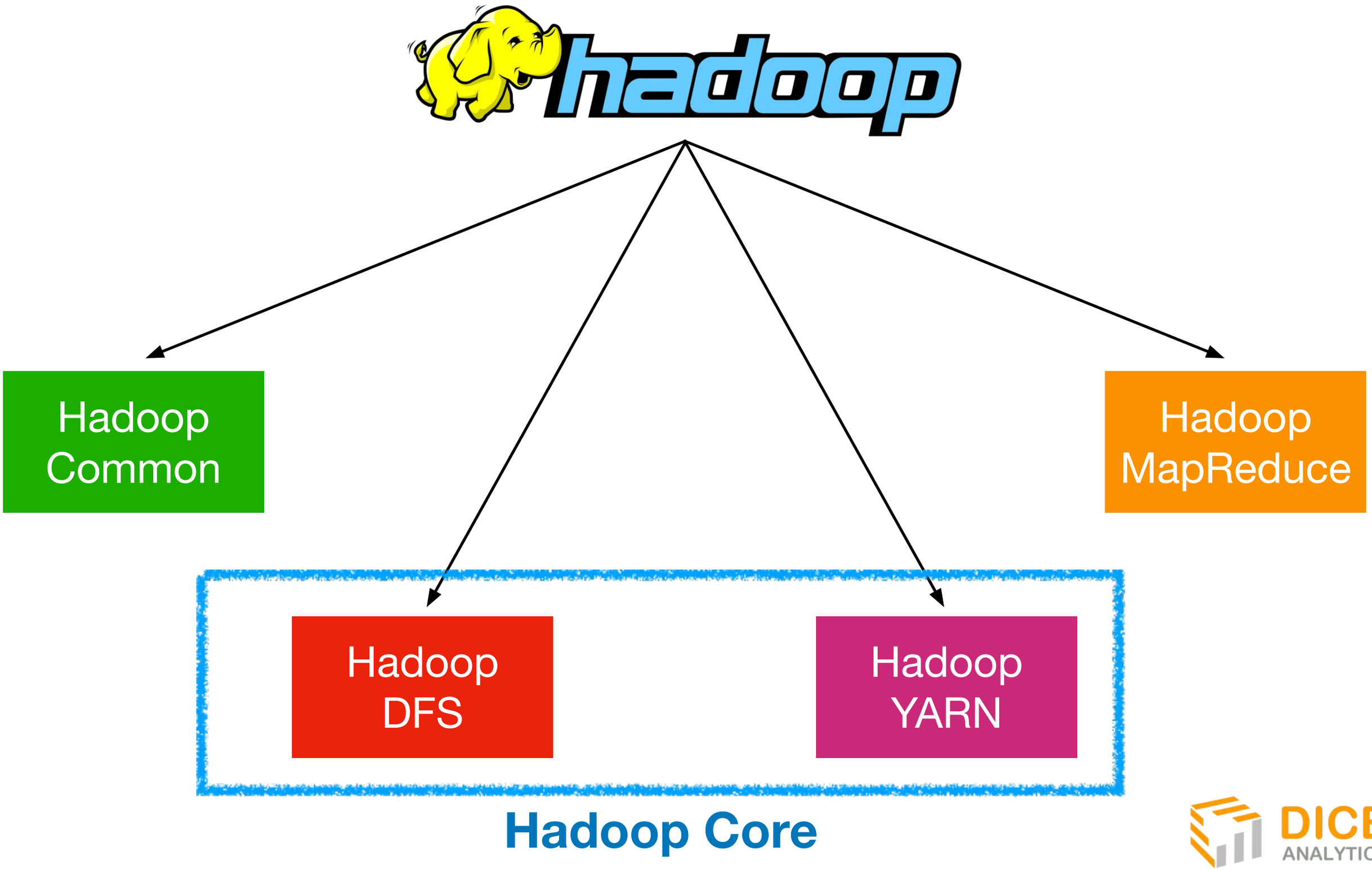
DICE ANALYTICS



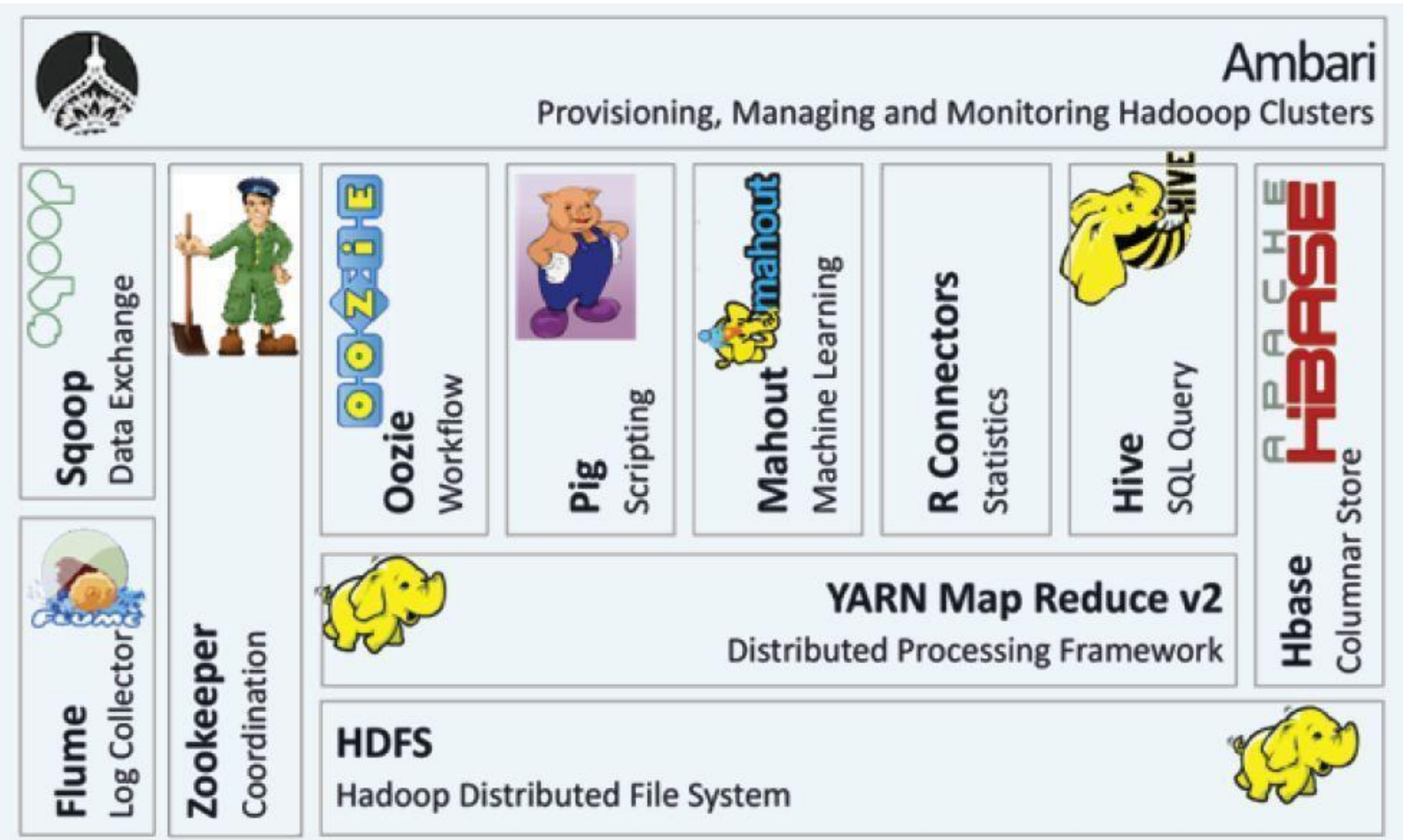



DICE
ANALYTICS

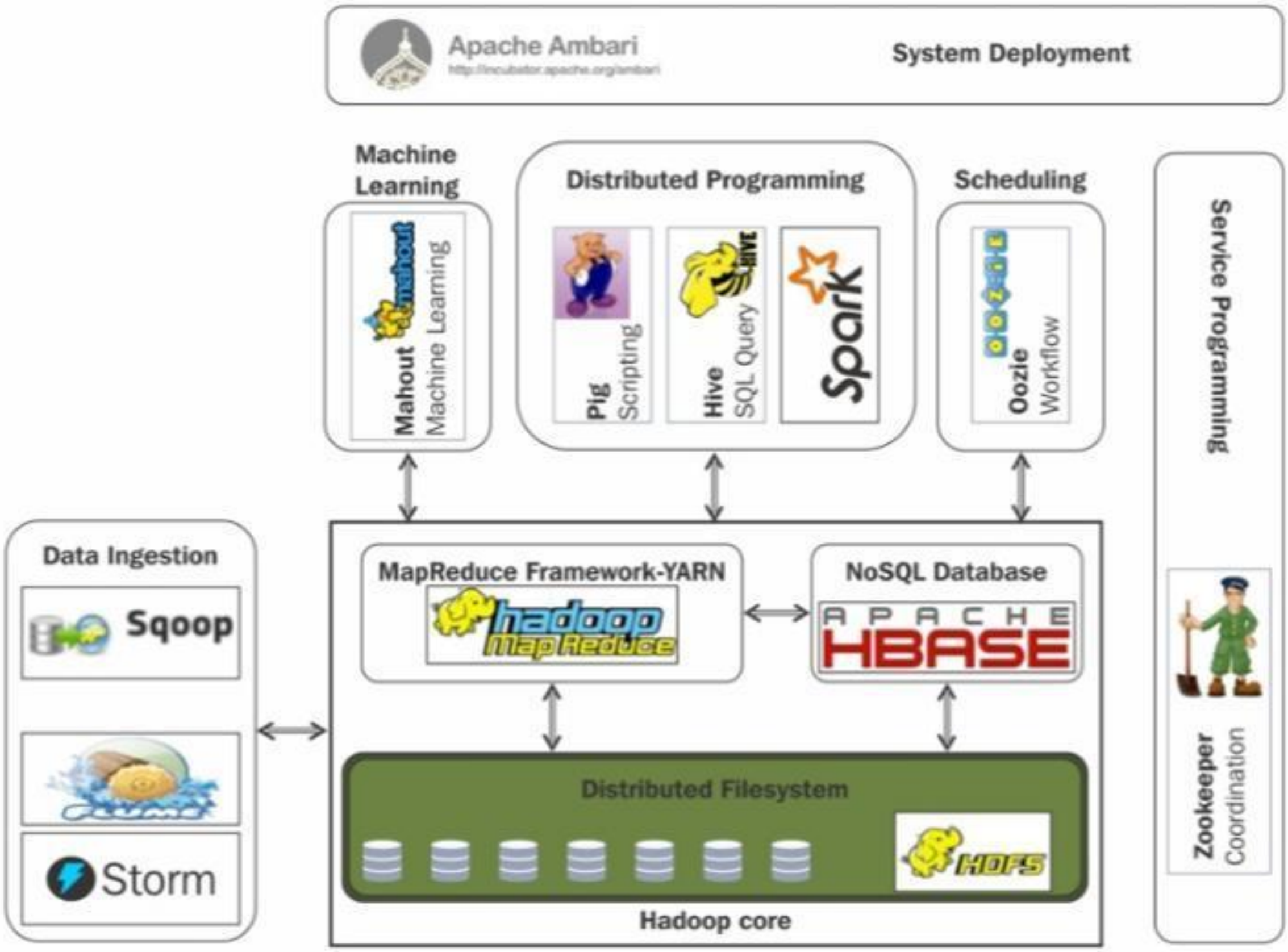
# Hadoop Composition



# Basic Hadoop Stack



# Basic Hadoop Stack



# Basic Hadoop Stack

## Data Management Frameworks


| | |
|---|---|
| HDFS | Hadoop Distributed File System.<br>A Java-based, distributed file system that provides scalable, reliable, high-throughput access to application data stored across commodity servers. |


| | |
|---|---|
| YARN | Yet Another Resource Negotiator.<br>A framework for cluster resource management and job scheduling. |

# Basic Hadoop Stack

## Operations Frameworks


| | |
|---|---|
| Ambari | A web-based framework for provisioning, managing and monitoring Hadoop Clusters. |


| | |
|---|---|
| Zookeeper | A high-performance coordination service for distributed applications. |


| | |
|---|---|
| Cloudbreak | A tool for provisioning and managing Hadoop Clusters in the cloud. |


| | |
|---|---|
| Oozie | A server-based workflow engine used to execute Hadoop Jobs |

# Basic Hadoop Stack

## Data Access Frameworks


| | |
|---|---|
| Pig | A high-level platform for extracting, transforming, analyzing large datasets. |


| | |
|---|---|
| Hive | A data warehouse infrastructure that supports ad hoc SQL queries. |

HCatalog

| | |
|---|---|
| HCatalog | A table information, schema and metadata management layer supporting Hive, Pig, MapReduce, and Tez Processing. |


| | |
|---|---|
| Cascading | Application development framework for building data applications, abstracting details of complex MapReduce programing. |


| | |
|---|---|
| HBase | A scalable distributed NoSQL database that supports structured data storage for large tables. |

# Basic Hadoop Stack

## Data Access Frameworks


| | |
|---|---|
| Phoenix | A client-side SQL layer over HBase that provides low latency access to HBase data. |


| | |
|---|---|
| Accumulo | A low latency, large table data storage and retrieval system with cell-level security. |


| | |
|---|---|
| Storm | A distributed computation system for processing continuous stream of real-time data. |


| | |
|---|---|
| Solr | A distributed search platform capable of indexing petabytes of data. |


| | |
|---|---|
| Spark | A fast, general purpose processing engine used to build and run sophisticated SQL, streaming, machine learning or graphics. |

# Basic Hadoop Stack

## Governance and Integration Frameworks


| | |
|---|---|
| Falcon | A data governance tool providing workflow orchestration, data lifecycle management, and data replication services. |

WebHDFS

| | |
|---|---|
| WebHDFS | A REST API that uses standard HTTP verbs to access, operate, manage HDFS. |

HDFS NFS Gateway

| | |
|---|---|
| HDFS NFS Gateway | A gateway that enables access to HDFS as an NFS mounted file system. |


| | |
|---|---|
| Flume | A distributed, reliable and highly available service that efficiently collects, aggregates and moves streaming data. |

# Basic Hadoop Stack

## Governance and Integration Frameworks

| Sqoop | A set of tools for importing and exporting data between Hadoop and RDBM systems. |
|---|---|


| Kafka | A fast, scalable, durable, and faut-tolerant publish-subscribe messaging system. |
|---|---|


| Atlas | A scalable and extensible set of core governance services enabling enterprises to meet compliance and data integration requirements. |
|---|---|

# Basic Hadoop Stack

## Security Frameworks

| | |
|---|---|
| HDFS | A storage management service providing file and directory permissions, even more granular file and directory access control lists, and transparent data encryption. |
| YARN | A resource management service with access control lists controlling access to compute resources and YARN administrative functions. |
| Hive | A data warehouse infrastructure service providing granular access controls to table columns and rows. |

# Basic Hadoop Stack

## Security Frameworks

| | | |
|---|---|---|
| APACHE FALCON | Falcon | A data governance tool providing access control lists that limit who may submit Hadoop Jobs. |
| APACHE KNOX | Knox | A gateway providing perimeter security to a Hadoop Cluster. |
| Apache Ranger | Ranger | A centralized security framework offering fine-grained policy controls for HDFS, Hive, Hbase, Knox, Storm, Kafka and Solr |

# Hadoop as +1 Architecture

- Though it has the potential to replace all others, it can also be used to complement existing systems if they can't be removed due to any constraints.