BIG DATA (TFC303)

Pertemuan 2 – *Big Data Stack Preparation*

ALIFIA REVAN PRANANDA

Department of Information Technology
Faculty of Engineering
Universitas Tidar

BIG DATA TECHNOLOGY

The effective use of big data cannot only **deliver** substantial top and bottom-line profits, but also **improve** the performance of existing function, and also **create** opportunity for growth and expansion.

In the organization, big data is also **efficient** to handle data management that consequent loss of business revenue. Even though Big Data has effectively increase the productivity, it still **remain challenges**.

The challenge is the strain that it put on their existing IT infrastructure due to **the huge data influx** and thereby slowing the system.

Hence, more robust architecture that able to handle the high volume and dynamic nature of big data is needed.

PROMISES TO BE TRANSFORMATIVE

BIG DATA TECHNOLOGY

To utilize **more robust architecture** of big data, we have to examine the following materials:

- ✓ Different architectural considerations associated with Big Data
- ✓ Big Data technology stack and its components

PROMISES TO BE TRANSFORMATIVE

BIG DATA TECHNOLOGY

✓ **Different architectural considerations associated with Big Data**

The architectural foundation of Big Data must take into account:

1. The functional and infrastructure requirement

→ This will encourage the development of the design principle.
It is extremely important for creating a strong environment and conducive for big data.

→ The architecture design must support the following functional requirements:

- Capture data
- Organize the data
- Integrate the data
- Analyse data
- Act on the results of the analysis
- Support good computational power & high speed computation
- Support the storage of huge volumes of data
- Have the right amount of redundancy to protect from any unexpected latency & downtime
- Including of consideration for infrastructure software, operational software, management software, well-defined API, & software development tools

BIG DATA TECHNOLOGY

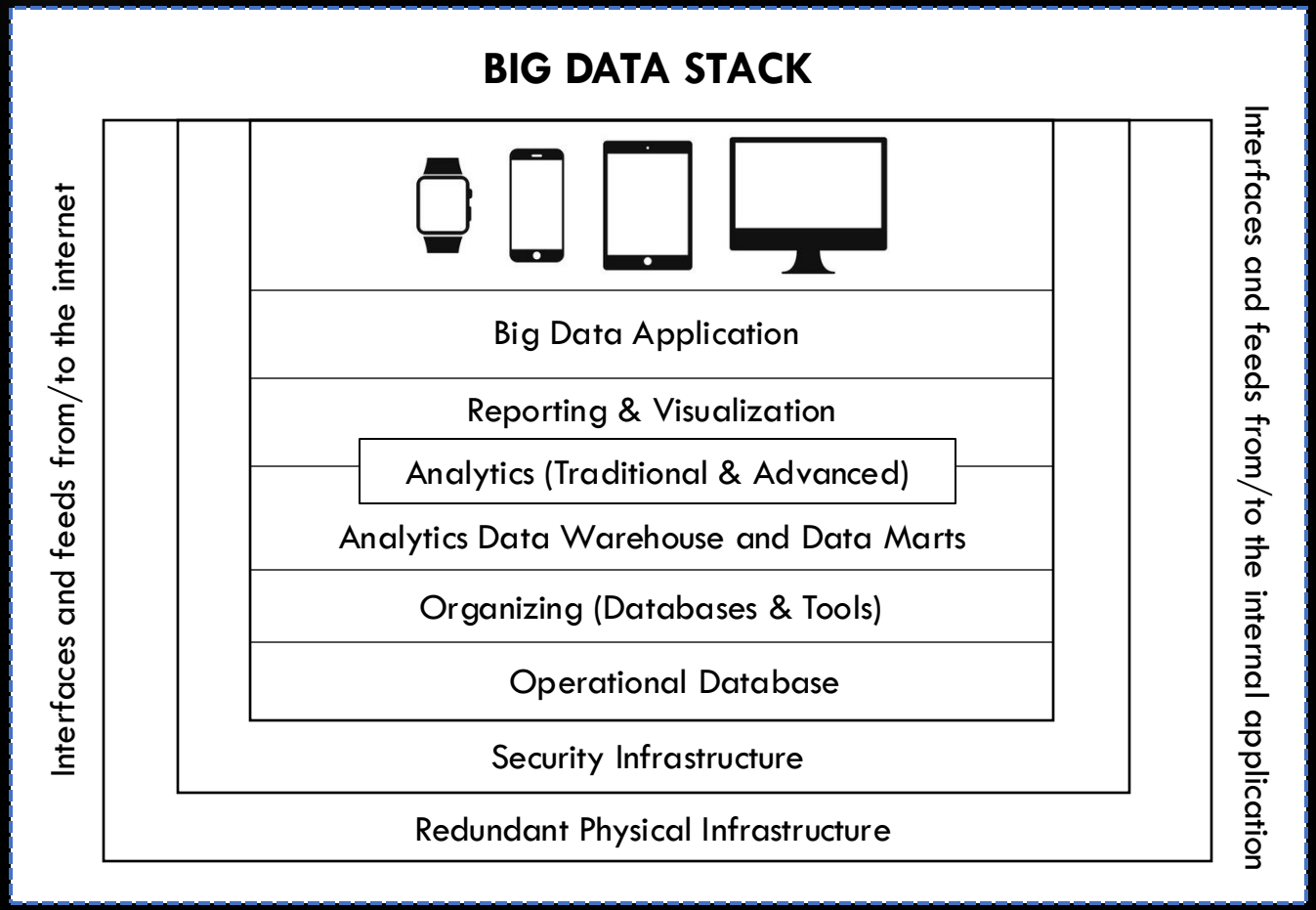
Thus, Big Data management architecture **must include variety of services** that enable companies to make use of a myriad data sources in fast and effective manner.

A framework of the Big Data technologies that can address the functional requirements for Big Data projects is referred to as a **TECHNOLOGY STACK**.

Let us understand **THE TECHNOLOGY STACK** in detail.

BIG DATA TECHNOLOGY STACK

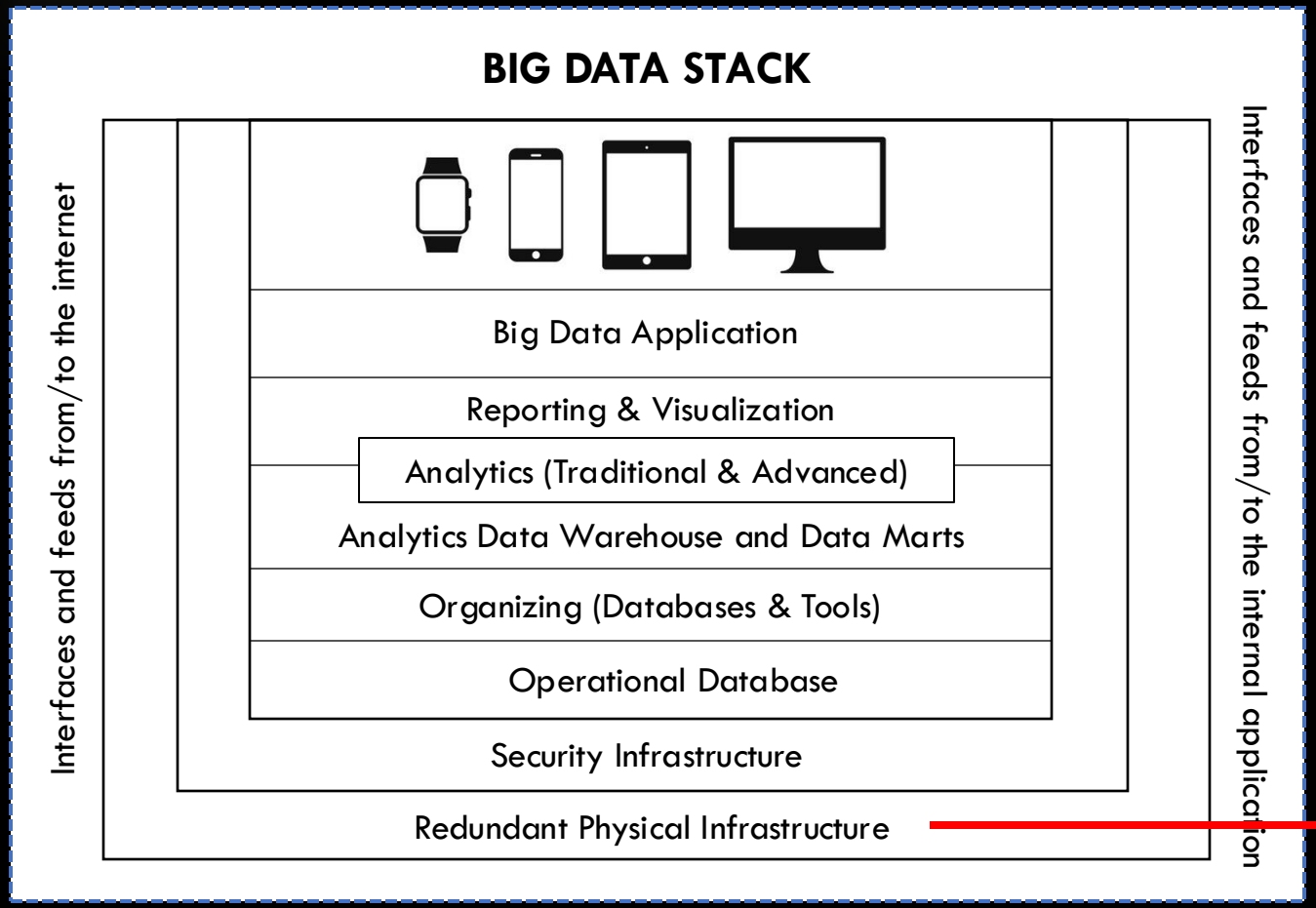
Big Data Technology Stack is a comprehensive stack that has several components addressing specific function of managing Big Data. There are 8 key layers of architecture:



These 8 layers represent the big data stack where each layer performs a different function. These layers work on a Big Data in tandem to produce the desired results.

BIG DATA TECHNOLOGY STACK

Big Data Technology Stack is a comprehensive stack that has several components addressing specific function of managing Big Data. There are 8 key layer of architecture:



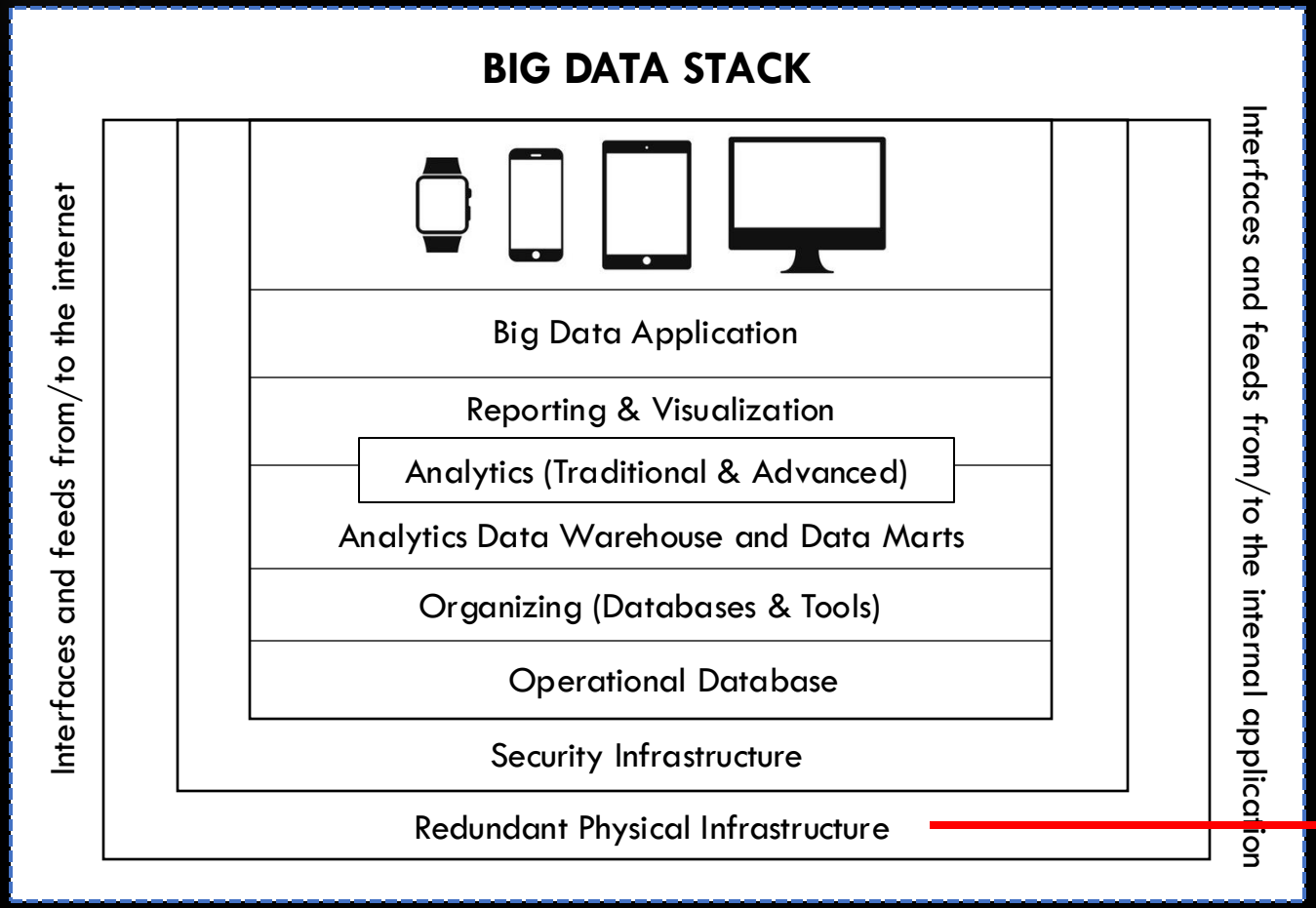
It comprises of the hardware and network elements. It is important to have some overarching principles to apply to your Big Data implementation. A prioritized of these principles should:

- **Performance**: refers to how responsive do you need the system to be.
- **Availability**: refers to do you need 100% uptime guarantee of services
- **Scalability**: refers to how big does your infrastructure need to be and how much computing power do you need.
- **Flexibility**: refers to how quickly can you add more resources to the infrastructure.
- **Cost**: what can you afford.

Networks, servers, and physical storage must be both resilient and redundant.

BIG DATA TECHNOLOGY STACK

Big Data Technology Stack is a comprehensive stack that has several components addressing specific function of managing Big Data. There are 8 key layer of architecture:



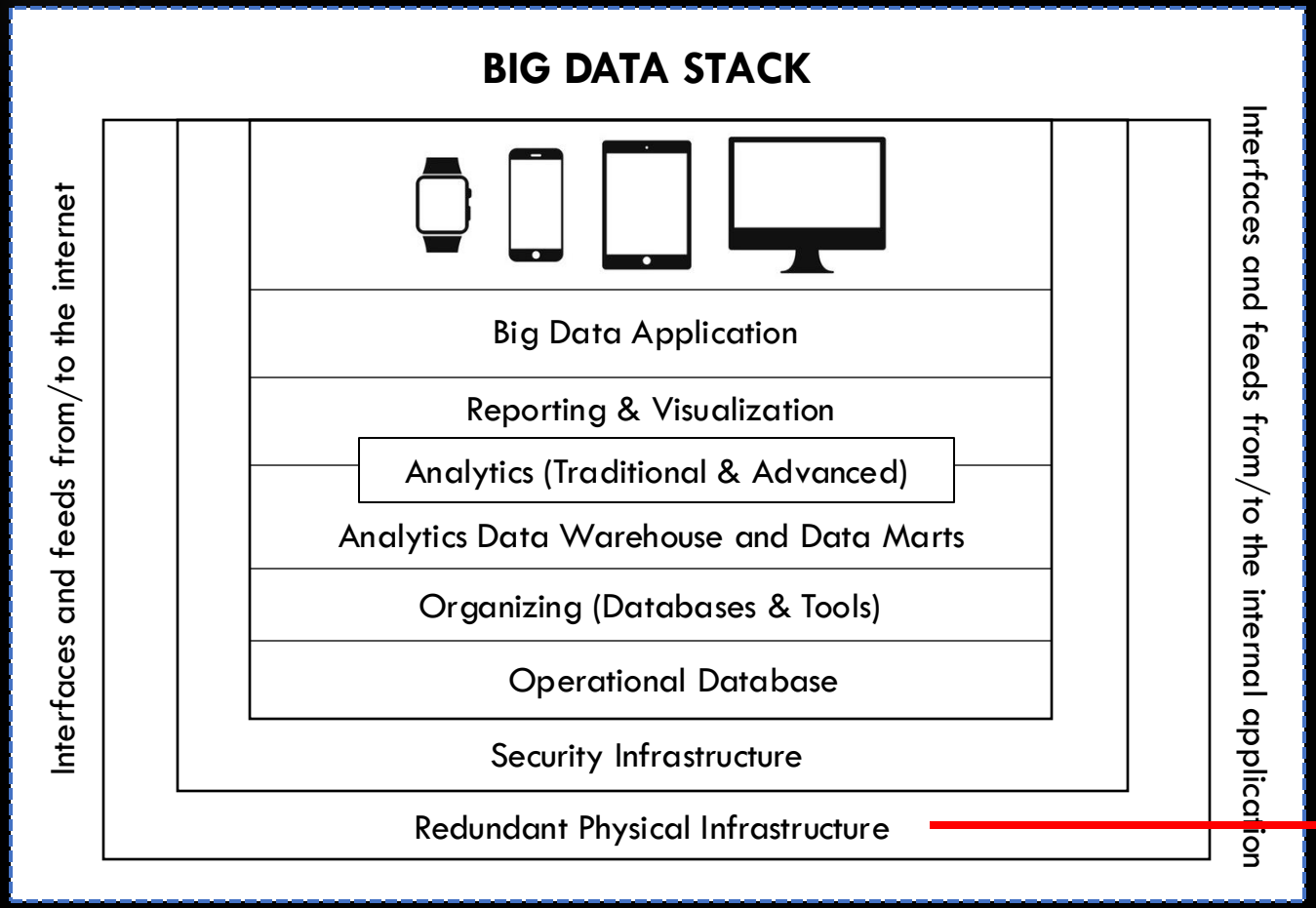
It comprises of three parts:

1. Physical redundant networks
   A business network should be redundant and must have the capacity to accommodate the anticipated volume and velocity of the inbound and outbound of Big Data.

   The physical redundant network should:
   - Be elastic
   - Offer monitoring capabilities to cope with the ebbs and flows of network traffic

BIG DATA TECHNOLOGY STACK

Big Data Technology Stack is a comprehensive stack that has several components addressing specific function of managing Big Data. There are 8 key layer of architecture:



It comprises of three parts:

2. Managing Hardware

   A business hardware which is made up of its storage and server assets must have sufficient speed and capacity to handle all expected big data capabilities.

   However to achieve it, network performance must also be up to the mark.

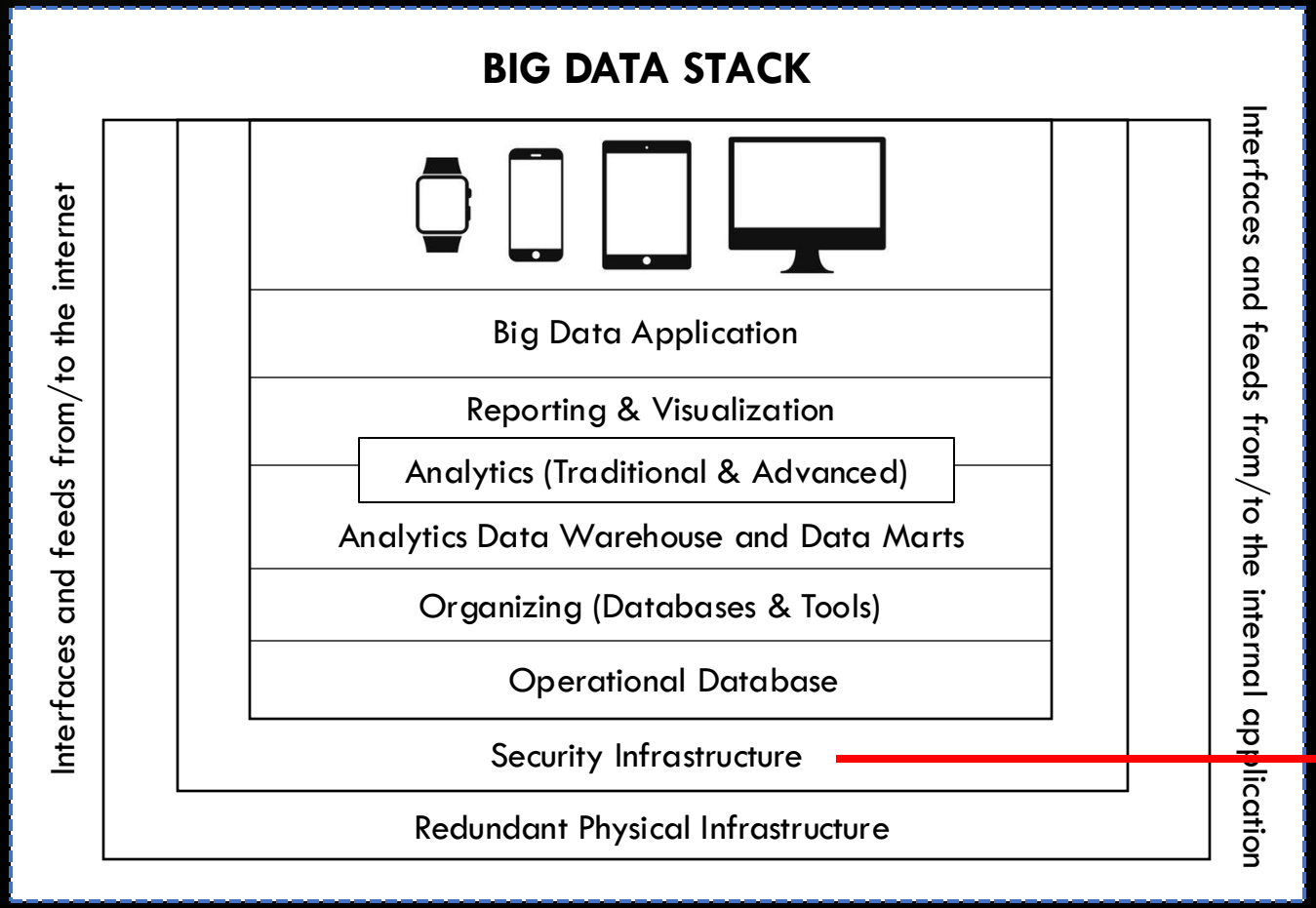
3. Infrastructure Operation

   This part aims to preparing the following works:

   - Optimum performance
   - Flexibility
   - Anticipate and prevent failures
   - Maintain integrity of data

BIG DATA TECHNOLOGY STACK

Big Data Technology Stack is a comprehensive stack that has several components addressing specific function of managing Big Data. There are 8 key layer of architecture:



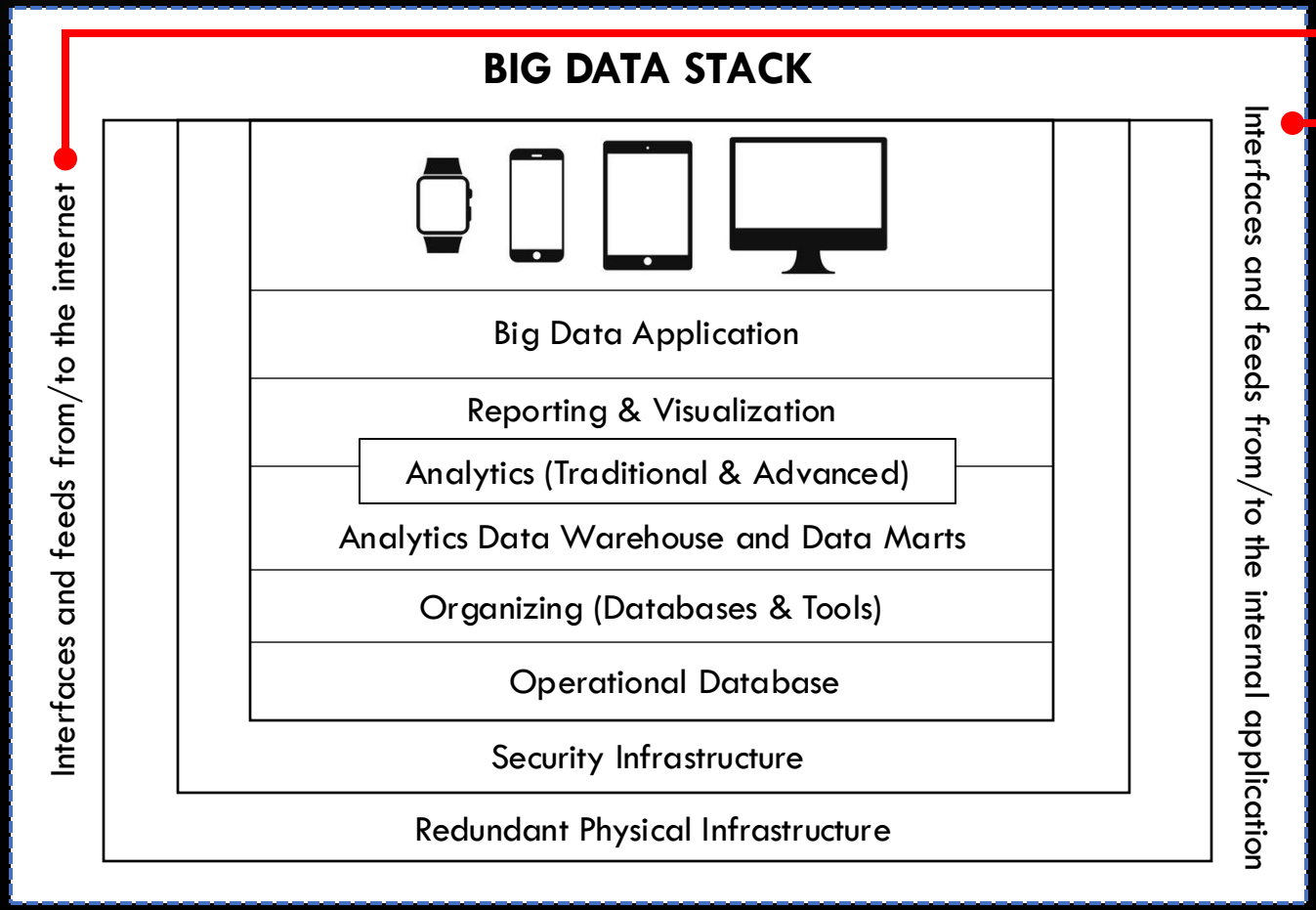
Security and privacy requirements for big data have to be closely aligned to specific needs.

Challenges:

- ✓ Data access & application access: these should be available only to who have a legitimate business need for examining and interacting with it
- ✓ Data encryption: it usually stresses the system resources. This problem can magnify in the case of big data.
- ✓ Threat detection: it stems from the inclusion of mobile devices and social network

BIG DATA TECHNOLOGY STACK

Big Data Technology Stack is a comprehensive stack that has several components addressing specific function of managing Big Data. There are 8 key layer of architecture:



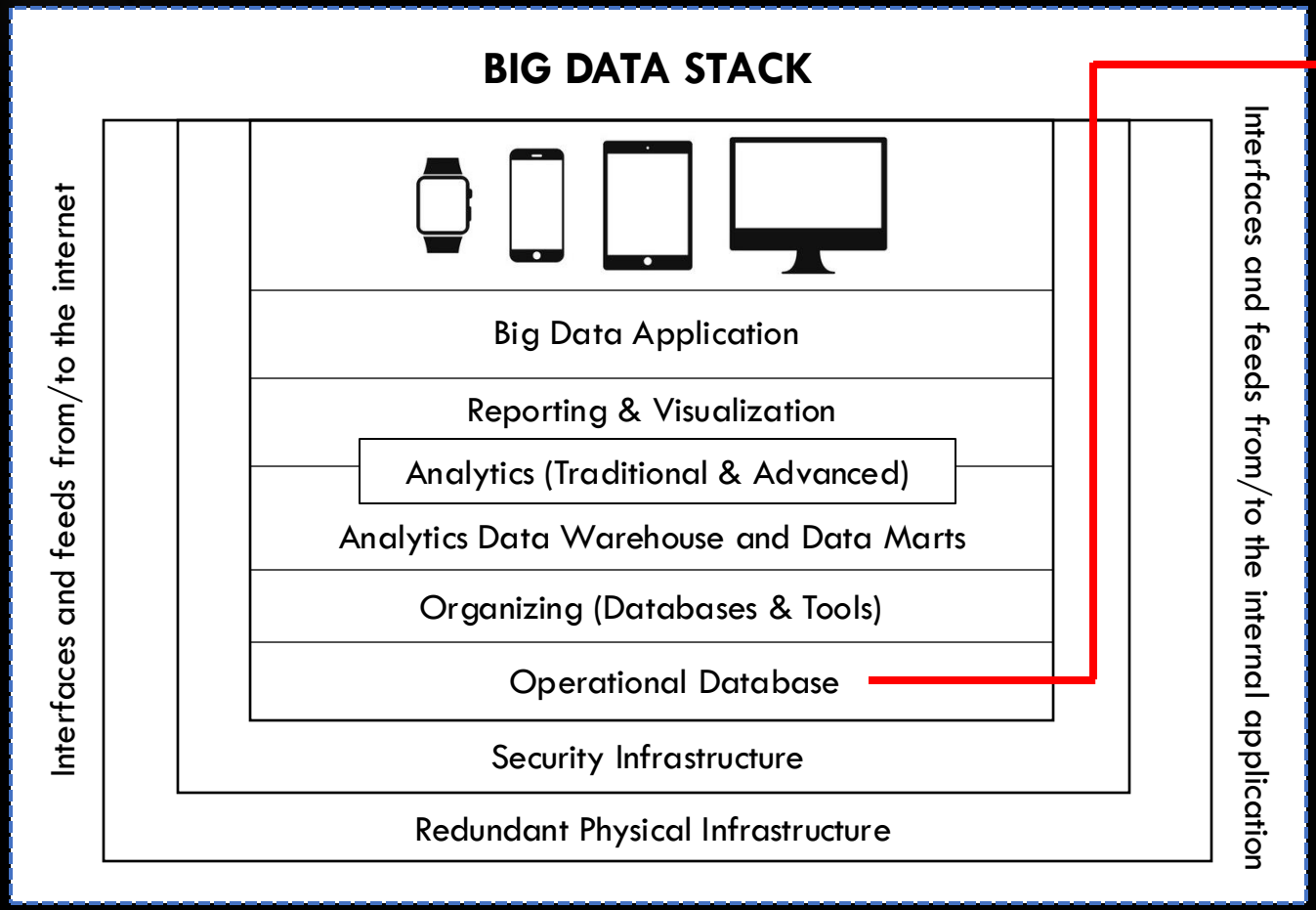
This layer manages the feed of data into and out of both internally managed data and data feeds from external sources.

Since big data relies on the fact that data from lots of sources are picked, this layer becomes critical for the big data solution.

Interfaces also exist at every level and between each layer of the stack. Without this layer, big data cannot happen.

BIG DATA TECHNOLOGY STACK

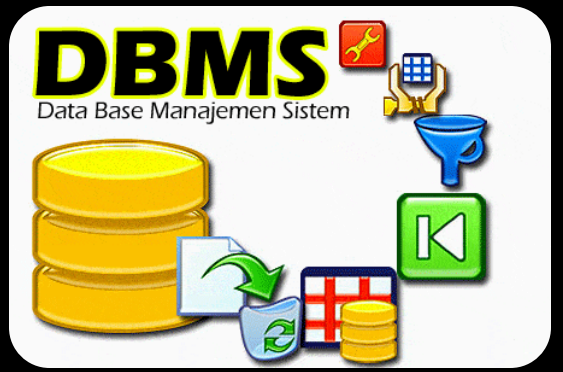
Big Data Technology Stack is a comprehensive stack that has several components addressing specific function of managing Big Data. There are 8 key layer of architecture:



The big data environment comprises of fast, scalable, and rock solid database engines that contain of collection of data element.

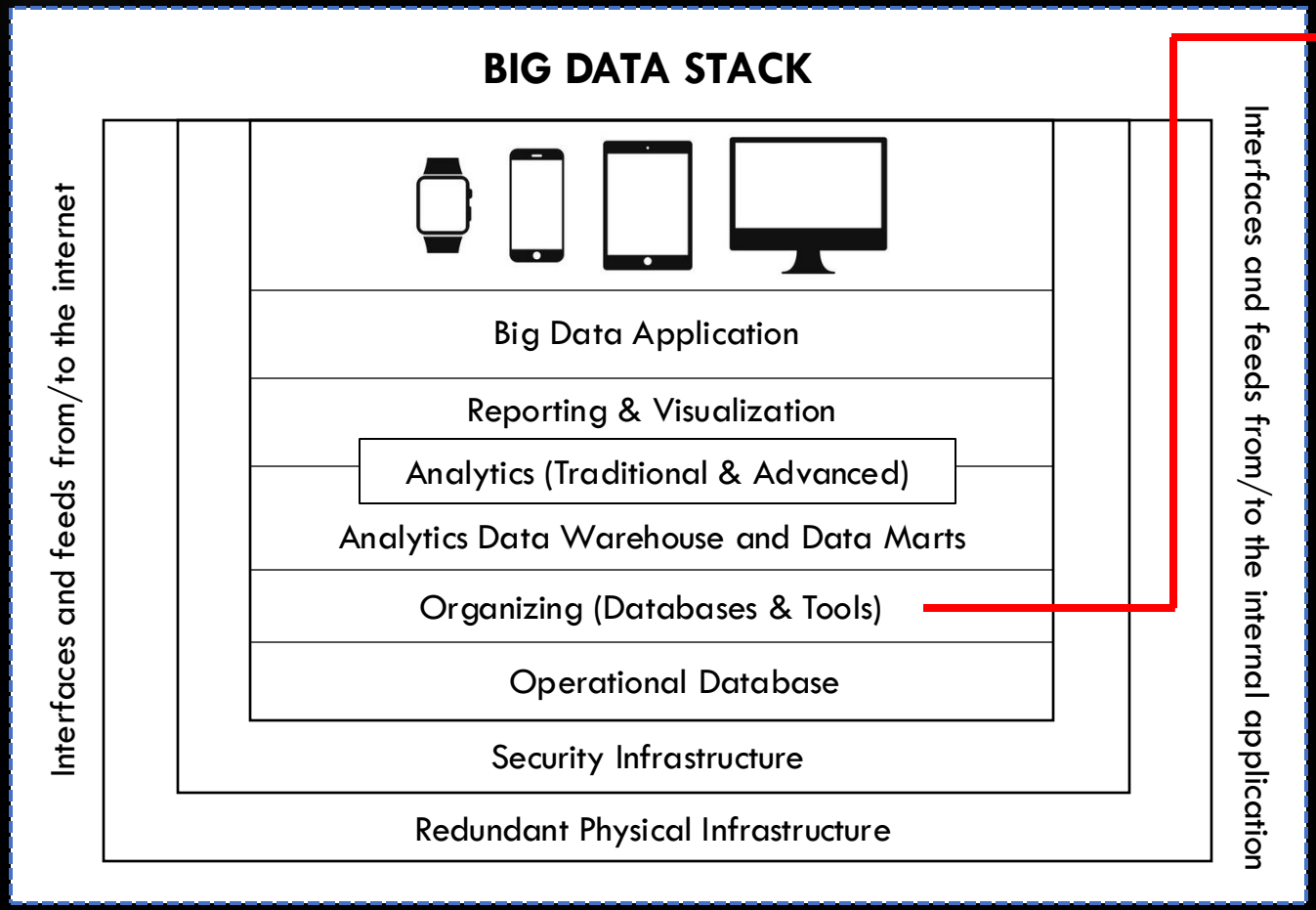
We have to understand about:

- ✓ Database engines
- ✓ Database languages



BIG DATA TECHNOLOGY STACK

Big Data Technology Stack is a comprehensive stack that has several components addressing specific function of managing Big Data. There are 8 key layer of architecture:



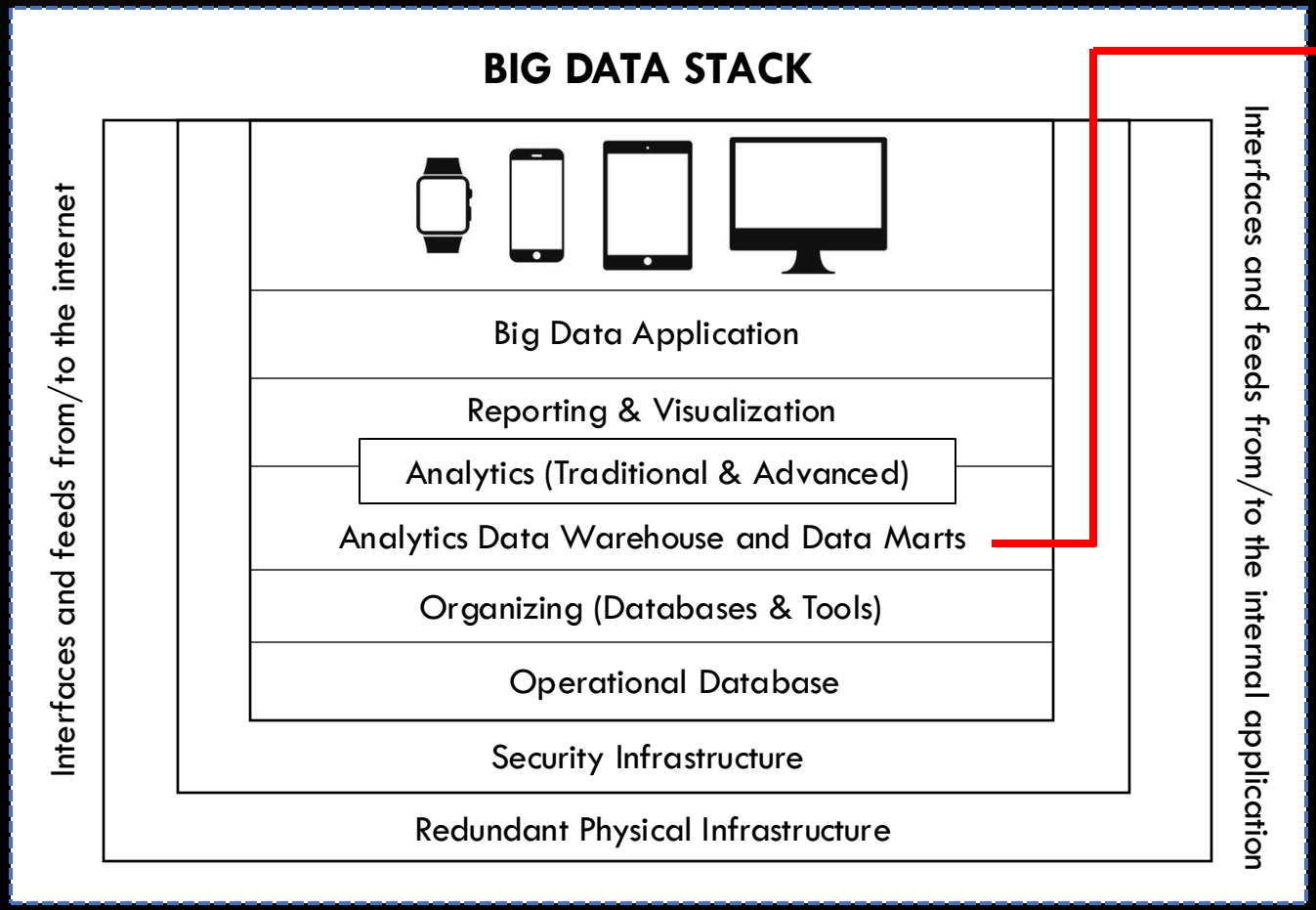
This layer comprises of an ecosystem of tools and technologies that can be used to gather and assemble for further processing.

The technology of this layer include of:

- A distributed file system
  - It is necessary to accommodate the decomposition of data stream.
  - It provides scales and storage capacity.
- Serialization services: it is necessary for persistent data storage and multi-language remote procedure calls.
- Coordination services: it is necessary for building distributed applications
- Extract, transform and load tools
- Workflow service for scheduling jobs

BIG DATA TECHNOLOGY STACK

Big Data Technology Stack is a comprehensive stack that has several components addressing specific function of managing Big Data. There are 8 key layer of architecture:

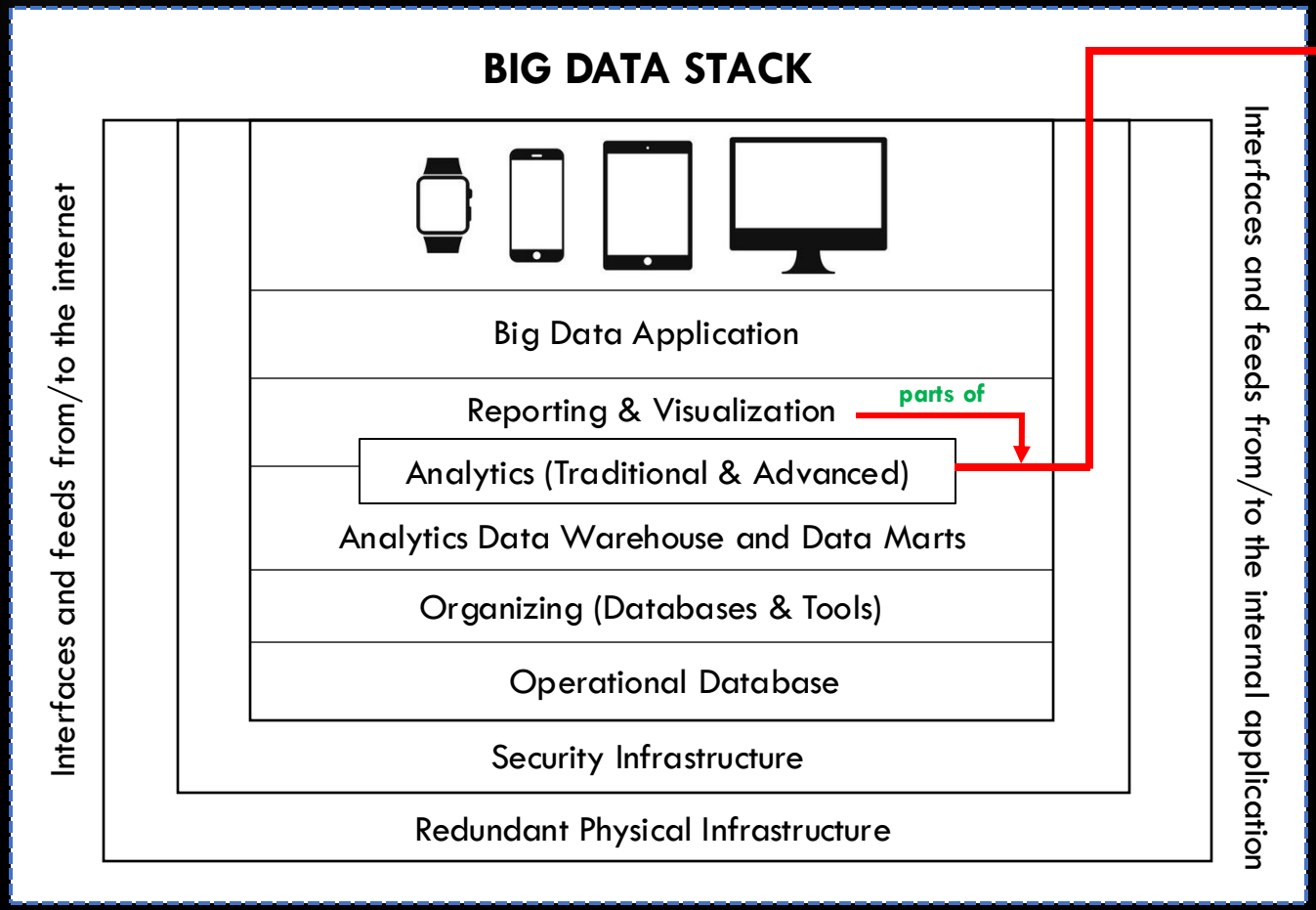


After sorting through the huge volumes of big data, it is necessary to extract the information to the end users in a user-friendly form.

These data warehouses and data marts provide compression multilevel partitioning and massively parallel processing architecture for preparing the data that reveals patters for analysis.

BIG DATA TECHNOLOGY STACK

Big Data Technology Stack is a comprehensive stack that has several components addressing specific function of managing Big Data. There are 8 key layer of architecture:



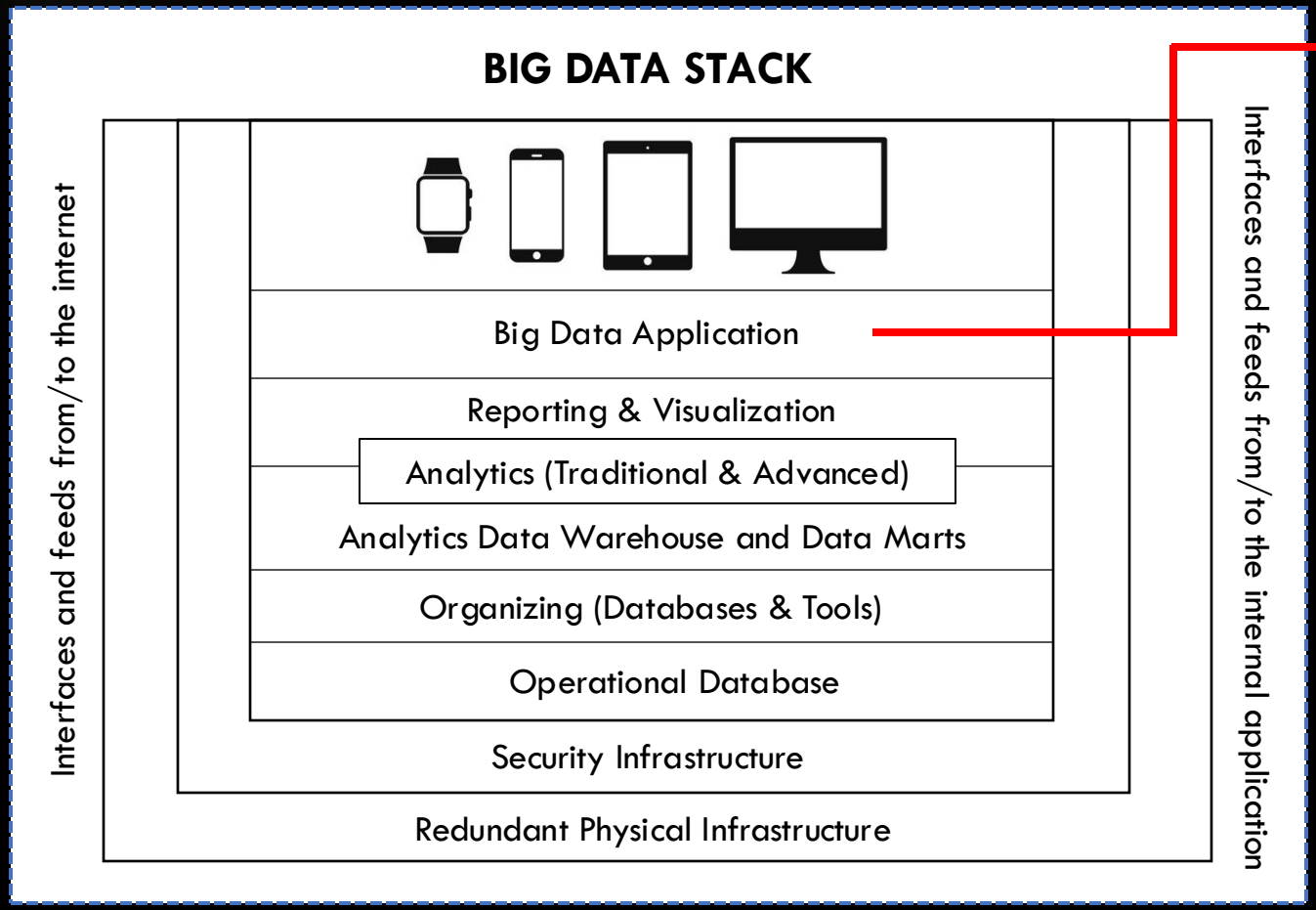
The existing analytics tools and techniques may not be able to work large amounts of potentially real-time and dissimilar data.

There are three classes of tools in this layer:

- Analytics and advanced analytics: tools that reach into the data warehouse and process the data for future use (e.g., predictive & sentiment analytics)
- Reporting and dashboard: provide user-friendly information from various resources.
- Visualization: allows business users to watch the changes of the data using different visualization techniques (e.g., mind map, heat map, infographic & connection diagrams)

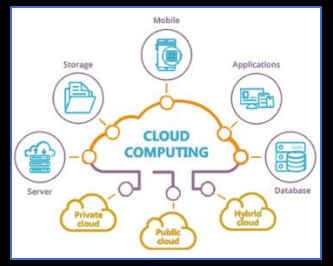
BIG DATA TECHNOLOGY STACK

Big Data Technology Stack is a comprehensive stack that has several components addressing specific function of managing Big Data. There are 8 key layer of architecture:



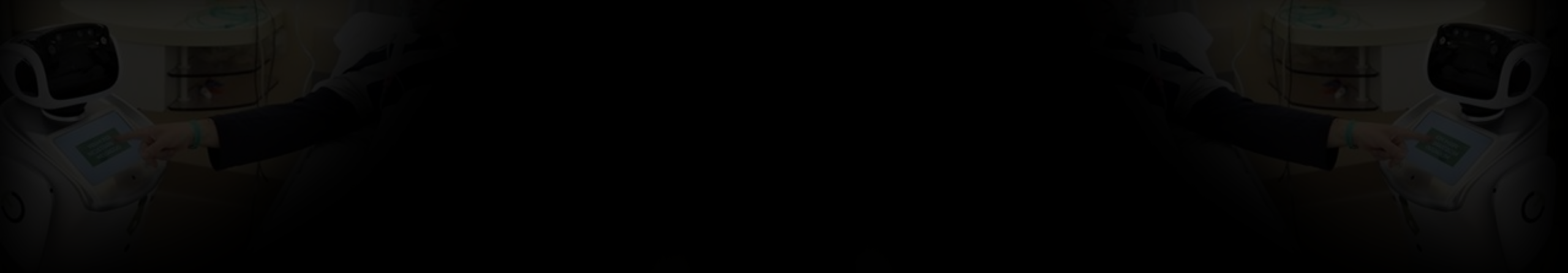
This application rely on huge volumes, velocity and variety of data to transform the behavior of a market. Big data application are either horizontal (they address problem that are common across industries) or vertical (they are intended to help solve an industry specific problem).

The creation of big data application will require structure, standard, rigor and well-defined API's.



Cloud Computing (example of big data application)

Hadoop (the most popular of Big Data technological framework)

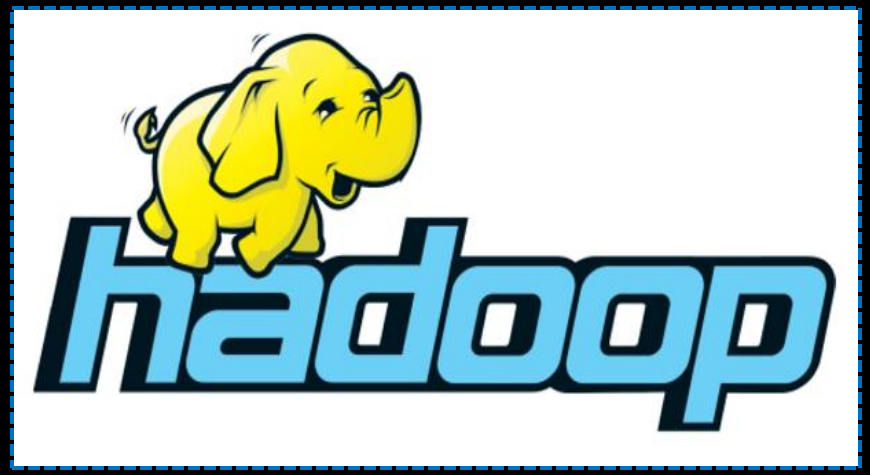


PREPARING BIG DATA STACK

BIG DATA STACK

The analytic process of big data should begin with the well-organized data. Big data stack technology offers effective solution in the data collection before we start to analyze them. Here are some big data stack framework that offer significant performance for our duties.

OPEN SOURCE FRAMEWORK



Hadoop
(the most popular of Big Data technological framework)

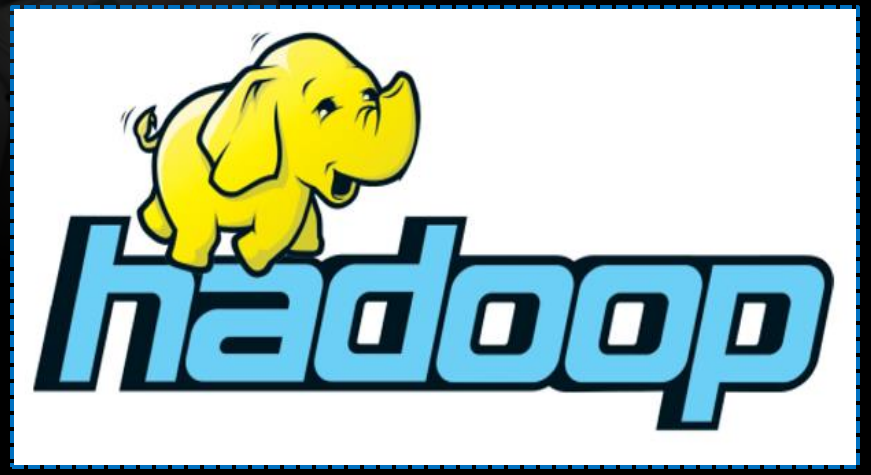
COMERCIAL VERSION

cloudera®

amazon web services

Microsoft Azure

HADOOP



Hadoop was born because of search engine like Yahoo and Google needed to find a way to understand the massive amounts collected by their engines. It offers:

- ✓ Flexibility
- ✓ Scalability
- ✓ Open source
- ✓ Fault tolerant
- ✓ Reallocates work automatically
- ✓ Real-time, costless, & cloud computing friendly.

Google

yahoo!

Hadoop allow us to manage and process the huge volume of data easily. Hadoop also manages different types of big data whether structure, unstructured or semi-structure.

HADOOP

The core component of Hadoop are:

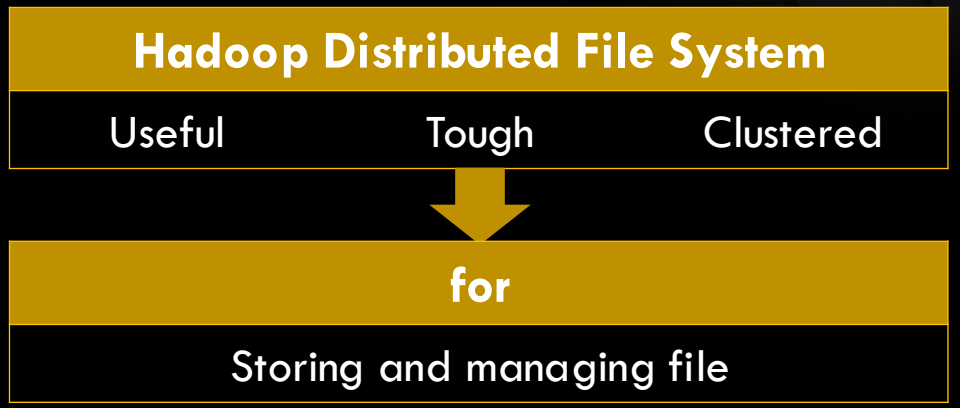
1. Hadoop Distributed File System (HDFS)
2. Hadoop MapReduce

These components provide the basic structures and services needed to support the key requirement of big data solution. However, just these two components are not enough to manage big data. Hadoop ecosystem therefore provides a collection of tools and techniques for the complete development and development of big data. Hadoop also provide the common technological framework for the different stakeholder such as developers database, administrators, and network managers who build big data solution.

The additional tools and technologies included in the Hadoop are:

- Hbase is the columnar store
- Sqoop is responsible for data exchange
- Hive manages the SQL queries
- Pig does the scripting
- etc.

Hadoop Distributed File System (HDFS)

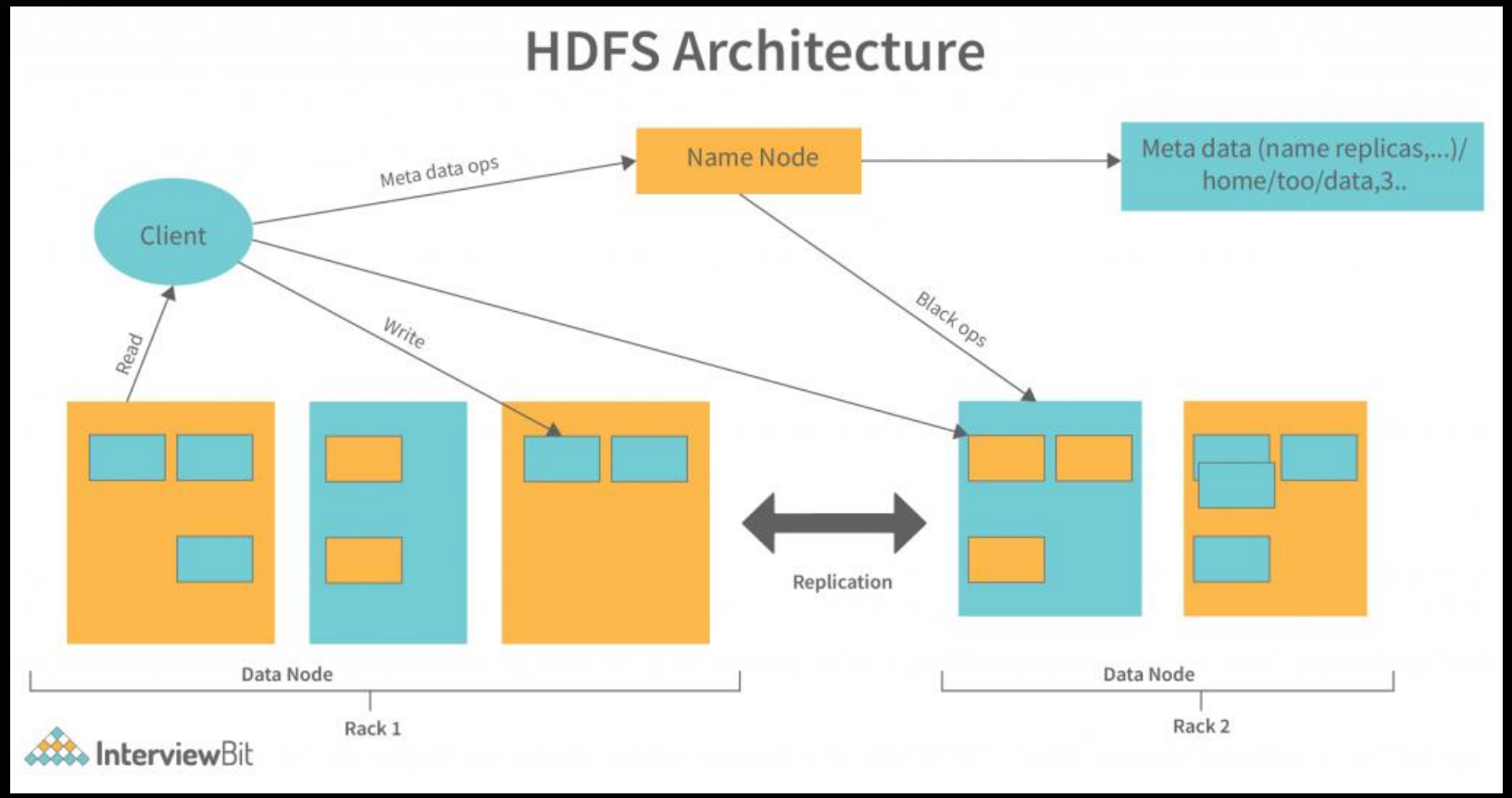


However, remember that HDFS is not the final destination for files. HDFS is rather a data services that offers a unique set of capabilities needed when the data volume and velocity are high.

HADOOP

How HDFS work?

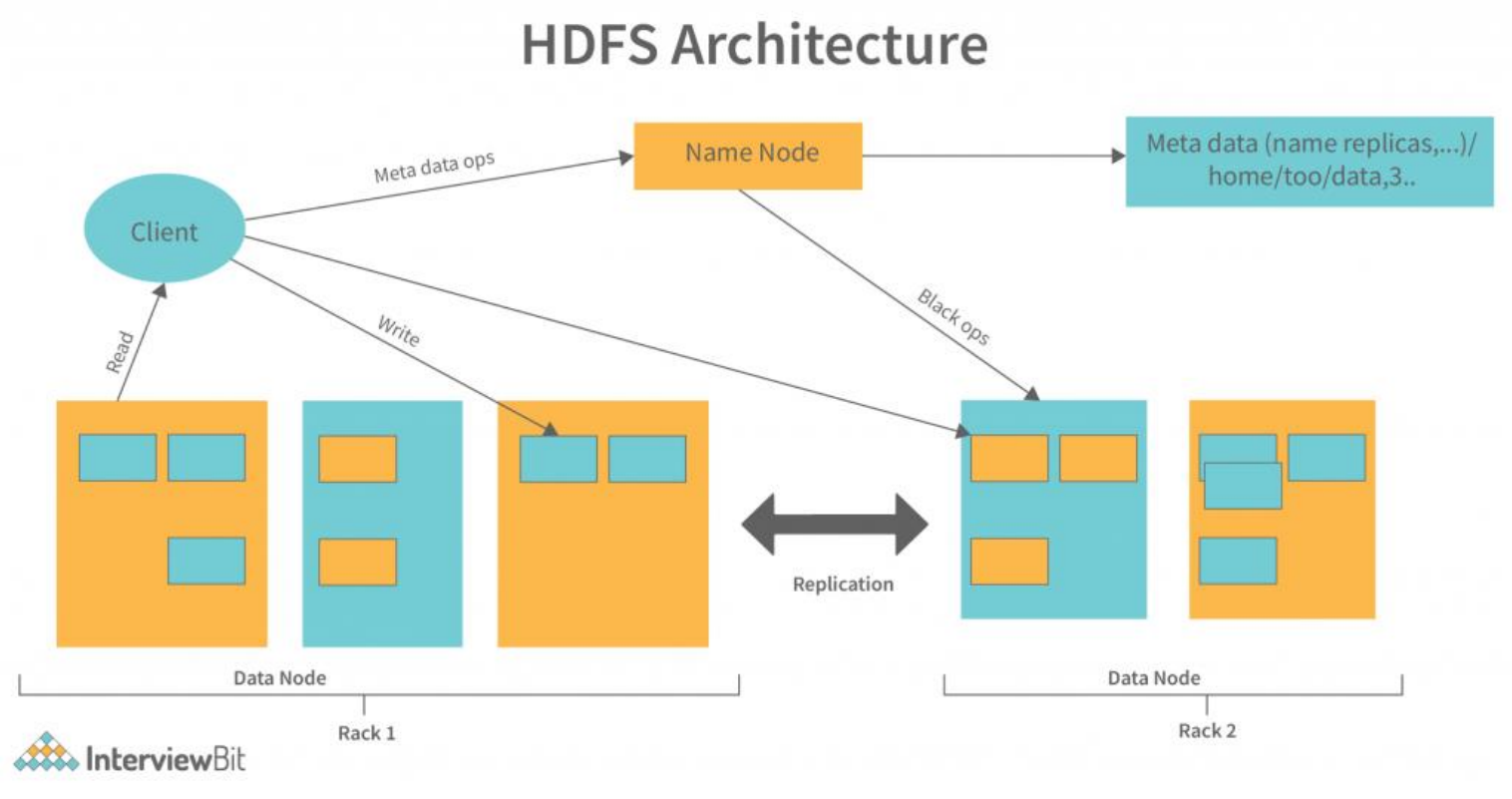
Distributed file system is a method of storing and accessing file based on a master-slave architecture. Files are stored on one or more central servers and can be accessed with proper authorization right by any number of remote clients in the network.



- ✓ HDFS follows a master-slave cluster design.
- ✓ Cluster is a networked set of nodes.
- ✓ Each HDFS cluster has a single master NameNode and several slave DataNodes.
- ✓ NameNodes : used for administrative function such as opening, closing and rename the files.
- ✓ DataNodes : perform the read and write function.

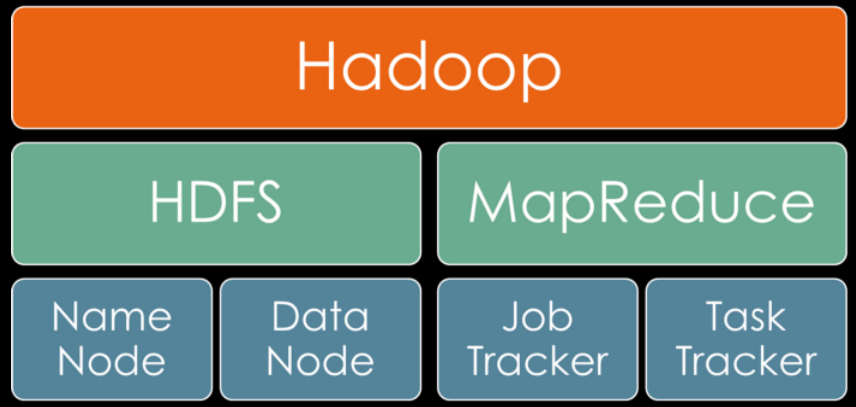
HADOOP

How HDFS work?



The way HDFS work is by breaking files into a collection of smaller block that are distributed among the data nodes in the HDFS cluster. They are managed by the NameNode.

All the DataNodes are collected into racks. Racks are physical collection of DataNodes in a single location. The NameNode should use the rack ID to keep track of all data nodes in the cluster. The NameNode can tracks the blocks on various DataNode that make up a complete file.



HADOOP

Advantages of Hadoop

- Break large amount of data into smaller blocks for better data handling
- Resilient: the blocks of data are replicated throughout the cluster to ensure data safety in case of server failures
- Implements data safety, data replication, data availability, and data resilience

Example:

A file is used to store contact numbers of people who live in New York. Because the file contains of huge amount of data, it is stored across an HDFS cluster.

Here the illustration of its servers.

SERVER-1 Contain of contact number of people whose last name begins with A	**SERVER-2** Contain of contact number of people whose last name begins with B
SERVER-4 Contain of contact number of people whose last name begins with D	**SERVER-3** Contain of contact number of people whose last name begins with C

A program collects blocks of file and information from each server to reconnect the original phonebook. In case one or few components of the server failed to respond, the HDFS replicates file blocks on to servers by default.

To provide the necessary data, this redundancy of data offers high availability and increase or decrease on the per file basis.

HADOOP

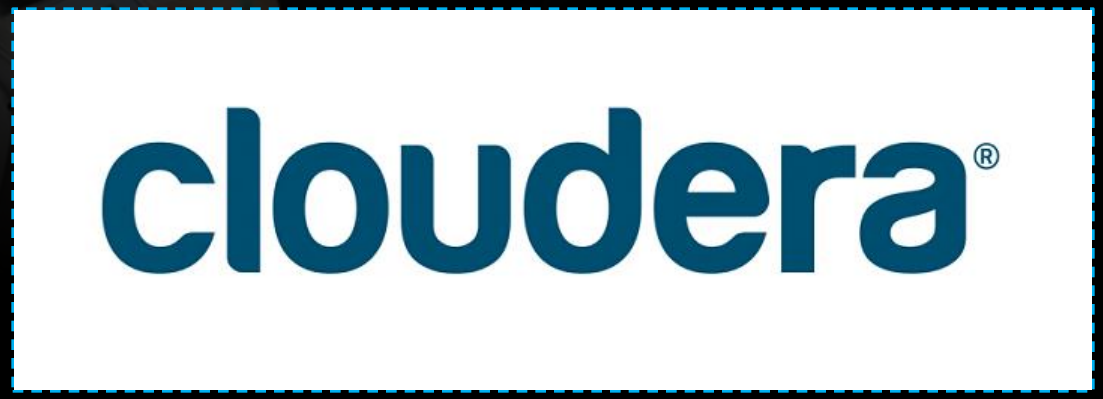
Disadvantages of Hadoop

- It is important to have a backup strategy in place. The cost of downtime can be extremely high, so it is important to keep things running smoothly. It is also recommended to have a security plan in place. If your company does not have a data backup plan, you are putting your company's data at risk.

- The chances are that the data in one location is vulnerable to hacking. Imagine the fear of losing valuable data when a disaster strikes. To protect data, backup data to a remote location. In the event of a disaster, the data can be quickly restored to its original location.

- This can be done manually or through a data migration process. Once the data is copied to the local environment, it can be accessed, analyzed, and used for any purpose.

CLOUDERA



As mentioned in the previous part, Cloudera is one of the commercially supported distribution of Hadoop. In 2008, Cloudera was founded by the smartest brains in Silicon Valley's leading companies, including Oracle, Facebook, Google, and Yahoo!.

Cloudera Enterprise is well fit for organizations that want to start up their own Enterprise Big Data Hub and perform data analysis on it. Cloudera Enterprise leverage open source Cloudera Distribution of Hadoop (CDH), which is one of the most deployed implementations of Hadoop in use today.

- The main four features provided by Cloudera are open source data platform, analytics, data management and predictive modelling.
- Cloudera Data Platform (CDP) have the best technologies of Hortonworks and Cloudera to serve as the first enterprise data cloud.
- CDP not only includes Data Hub service but also Data Warehouse and Machine Learning.
- It provided an integrated control plane to manage the data, infrastructure and analytic, CDP also support hybrid cloud or multi-cloud environment.
- CDP is fully open source distribution, vendor lock-in can be avoided.
- For pricing, Cloudera offers various options which includes annual subscriptions starting at $4,000 and $0.08 per hour for services used

AMAZONE WEB SERVICE (AWS)

Amazon as the leader in global online sales, has its own cloud computing, namely Amazon Web Services (AWS). In 2006, AWS was founded and begin to provide cloud computing to individuals and organizations. Amazon Web Services provides highly customizable services, in term of storage and service used. Furthermore, the cost is calculated based on the number of services selected. It is highly beneficial for both small and large organization.

- The cloud computing services provided by AWS can be divides into 19 categories, which are Compute, Storage, Database, Migration, Networking and Content Delivery, Developer Tools, Management Tools, Artificial Intelligence, Analytics, Security, Identity and Compliance, Mobile Services, Application Services, Messaging, Business Productivity, Desktop and App Streaming, Software, Internet of Things, Contact Center and Game Development.
- Amazon have strong investment in AI, they provide Amazon Machine Learning which capable for real-time predictions and contains of visualization tools and wizards.
- The biggest benefit AWS provides is pay per use, which make it not only suitable for large organizations, but also small and medium enterprise, or even individuals.

MICROSOFT AZURE

Microsoft Azure which was first released in 2010. Microsoft Azure enable users to run service on the cloud, or combine the cloud computing service with any infrastructure, data center or applications. Microsoft Azure provide a wide variety of services for different kinds of industry. They considered all kinds of the business needs and came out with various packages which sufficient to fulfill the needs of various kinds of industry.

- The cloud services provided by Microsoft Azure can be divided into 14 categories, which are Compute, Networking, Storage, Web + Mobile, Containers, Databases, Data + Analytics, AI + Cognitive Services, Internet of Things, Enterprise Integration, Security + Identity, Developer Tools, Monitoring + Management, and Microsoft Azure Stack.
- Microsoft Azure is benefit for individuals or organizations those are already using Microsoft software, for example, Windows and Office. By using Microsoft Azure which have the interface which are familiar by Microsoft service users, they can adopt it quickly and some of the services provided by Microsoft Azure is free for Microsoft service subscribers

HIGHLIGHTS OF THE DISTRIBUTIONS / SERVICES AND ITS COMPONENTS

Criteria	Cloudera	AWS	Azure
Price Flexibility	Multiple tier	Pay per use	Can get some free services if users are already Windows services subscriber
Hybrid Cloud Environment	Support	Does not support	Support
Loading large data	Strong	Weak	Moderate
Query on many dataset	Strong	Weak	Moderate

Cloudera have deployed Hadoop the most and hence have really strong performance in handling complex data. AWS are best known with their price flexibility, which suit for every user. Microsoft Azure are the most fit with Windows services subscribers.

HOW TO PREPARING BIG DATA STACK FRAMEWORK

MICROSOFT AZURE DESKTOP SET UP

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

Original software publication

AKFruitData: A dual software application for Azure Kinect c acquire and extract informative data in yield tests perform orchard environments

Juan Carlos Miranda *, Jordi Gené-Mola, Jaume Arnó, Eduard Gregorio

Research Group in AgroICT & Precision Agriculture, Department of Agricultural and Forest Engineering, Universitat de Lleid CERCA Center, Rovira Roure 191, 25198, Lleida, Catalonia, Spain

ARTICLE INFO

Article history:
Received 7 June 2022
Received in revised form 5 September 2022
Accepted 11 October 2022

Dataset link: https://github.com/GRAP-UdL-AT/ak_acquisition_system, https://github.com/GRAP-UdL-AT/ak_frame_extractor

Keywords:
RGB-D camera
Data acquisition
Data extraction
Fruit yield trials
Precision fructiculture

ABSTRACT

The emergence of low-cost 3D sensors, and particularly RGB in artificial intelligence, is currently driving the developme size measurement and yield estimation. However, as the p the availability of quality fruit datasets, the development in agricultural environments is essential. The AKFruitData facilitate use of the Azure Kinect RGB-D camera for testing i structure that addresses both the data acquisition and the da (AK_ACQS) allows different sensors to be activated simult Then, the extraction software (AK_FRAEX) allows videos ge be processed to create the datasets, making available col AKFruitData has been used by the authors to acquire an subsequent fruit yield estimation. Moreover, this software in the framework of precision agriculture, thus making it a v in fruit growing.

International Journal of Applied Earth Observations and Geoinformation 112 (2022) 102895

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag

Low-cost mobile mapping system solution for traffic sign segmentation using Azure Kinect

Zhouyan Qiu [a,b,*], Joaquín Martínez-Sánchez [a], Víctor Manuel Brea [c], Paula López [c], Pedro Arias [a]

[a] *CINTECX, Universidade de Vigo, Applied Geotechnology Group, Vigo 36310, Spain*
[b] *ICT & Innovation Department, Ingeniería Insitu, Vigo 36310, Spain*
[c] *Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela 15782, Spain*

ARTICLE INFO

Keywords:
Time of Flight camera
Traffic sign detection and segmentation
Multi sensor system
Mobile mapping system

ABSTRACT

The mobile mapping system (MMS) could become the foundation of digital twins and 3D modeling, and is widely applicable in a variety of fields, such as infrastructure management, intelligent transportation systems, and smart cities. However, data collected by MMS is extensive and complex, making data processing difficult. We present a novel method for segmenting urban assets (specifically in this case study traffic signs) with a lower-cost Azure Kinect and automatic data processing workflows. First, it was necessary to verify the reliability of this approach using the Time of Flight (ToF) camera from Azure Kinect to detect road signs outdoors. Using the data generated by the ToF camera, we then extracted the Region of Interest (ROI) quickly and efficiently. After transforming the ROI to the RGB image, we obtained the traffic sign area through a hybrid color-shape based method. In addition, we calculated the distance between the traffic sign and Azure Kinect based on the depth image. The Coefficient of Variation c_v averaged 1.1%. It is thus evident that Azure Kinect is reliable for outdoor traffic sign segmentation. Our algorithm has been compared with deep learning algorithms. According to our analysis, our algorithm has an accuracy of 0.8216, while the accuracy of deep learning is 0.7466, which indicates that our solution is more flexible and cost-effective.

EXAMPLE CASES

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 176 (2020) 3710–3717

Procedia Computer Science

www.elsevier.com/locate/procedia

24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Usage visualisation for the AWS services

Lettisia Catherine George, Yanan Guo, Denis Stepanov, Vikas Kumar Reddy Peri, Roshan Lakmal Elvitigala, Maria Spichkova*

School of Science, RMIT University, Melbourne, Australia

Abstract

The goal of this work is to elaborate a lightweight solution to manage multiple Amazon Web Services (AWS) accounts in one place. Currently, no such system is available. In this paper, we present a system that solves this task and allows to visualise the usage of AWS services across a number of customer's accounts. This system can be also used to enable optimisation cost and performance of the AWS services across the customer's accounts.

Keywords: Software Engineering, Usability, Visualisation, AWS

Computers & Industrial Engineering
Volume 170, August 2022, 108331

AWS IoT analytics platform for microgrid operation management

Carlos Antonio Marino [a b 1], Flavia Chinelato [a b], Mohammad Marufuzzaman [c]

Show more

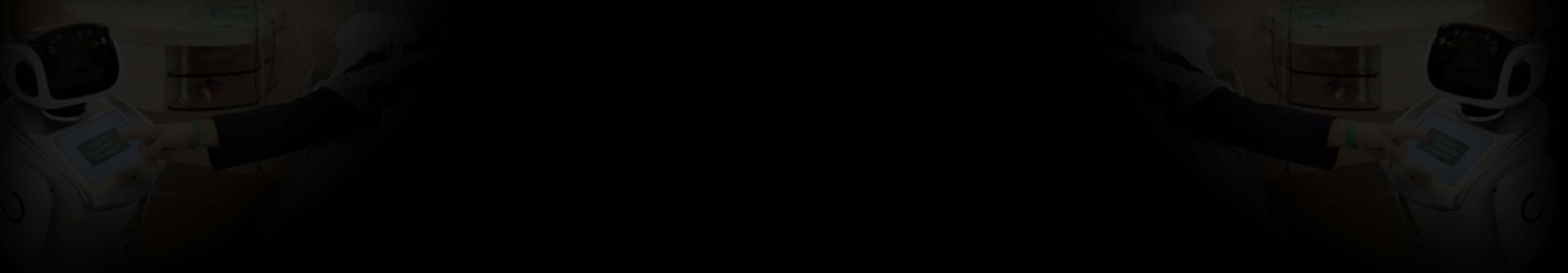
+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.cie.2022.108331 Get rights and content

Abstract

Microgrid (MG) represents a promising opportunity for integrating renewable energy systems with the electric power grid. However, numerous complexities need to be addressed in the process. The electrical grid is complex, vulnerable, and centralized. Thus, the integration is challenging owing to the stochastic nature of renewable energy generation, which affects the possibility of reliable forecasting. The wastage due to poor estimation of clean energy generation discourages new investments in this area. However, recent advancements in big data technologies enable processing a large amount of data captured from multiple sources in real-time. It opens the possibility of improving the operational optimization of MGs and the performance of forecasting models. The overall MG problem is formulated using a two-stage stochastic mixed-integer linear programming problem with recourse. Amazon Web Services (AWS) IoT analytics platform inputs data in real-time and runs a sophisticated wind generation forecast analysis. The stochastic model is solved using the Sample Average Approximation (SAA) algorithm. The innovative methodology leads to significant

EXAMPLE CASES



ANY QUESTION ?

TUGAS

1. Cari paper/jurnal yang memanfaatkan teknologi Big Data menggunakan Cloudera/AWS/Azure
2. Analisis bagaimana pemanfaatan Big Data pada penelitian tersebut

Dikerjakan secara individu. Deadline 9 September 2024 pukul 23:59 melalui ELITA. Hindari plagiarisme dan sertakan sumber referensi jika ada.