Introduction to Big Data

BIG DATA FUNDAMENTALS

Part II

Le Ngoc Thanh – Nguyen Ngoc Thao
{lnthanh, nnthao}@fit.hcmus.edu.vn

Outline

- Big data platform and Technologies
    - IBM Big data platform
- Digging into Big data technology
    - Big data technology stack
    - Big data analytics platforms and software

Big data platform

Comprehensive, enterprise-ready, integrated

Main tasks in Big data

- Tasks within the domain of Big Data often involve data mining as a prevalent method, yet at a larger scale.

Data aggregation

Data analysis

Data manipulation

Data visualization

Big data are multidisciplinary

- Technologies in Big data involves multidisciplinary studies.

Massively parallel processing databases

Distributed databases

Data mining grids

Strong internet

Scalable storage systems

Distributed filesystems

Cloud computing platforms

Economics, Statistics, Applied mathematics, and Computer Science, etc.

A Big data platform should offer

Comprehensive

Every dimension of Big data challenge is addressed.

Enterprise-ready

Features of performance, security, usability and reliability included.

Integrated

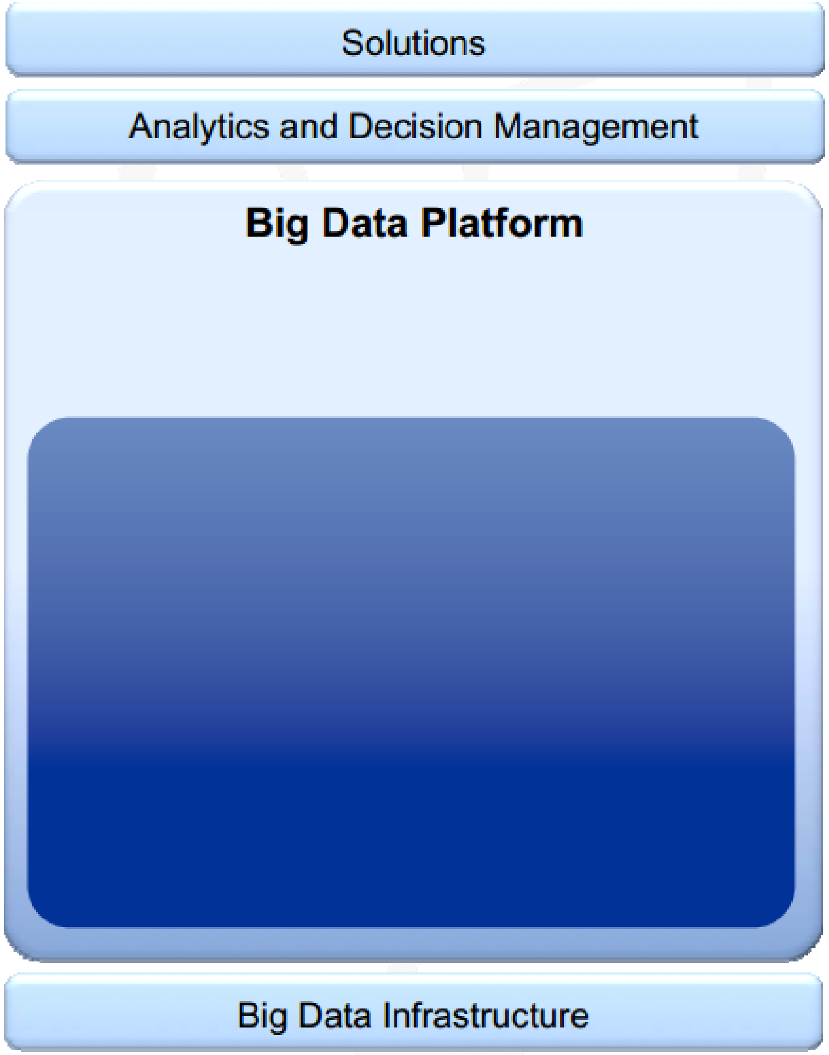
Big data technologies to enterprise should be simplified and accelerated

Integration with information supply chain, including databases, data warehouses, and BI applications.

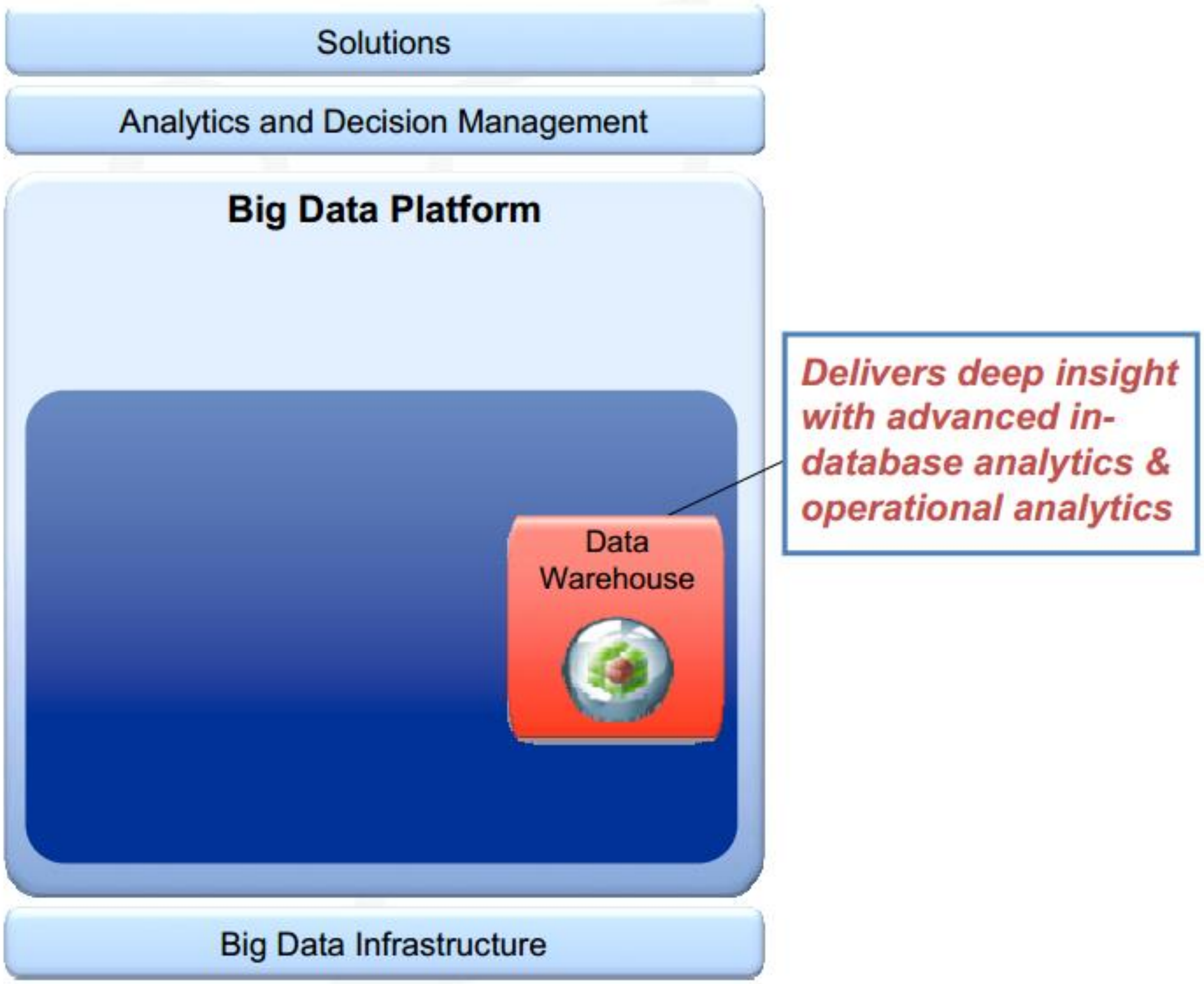
- Moreover, a Big Data platform should also offer
  - Open-source based, low latency reads/updates, ad-hoc queries, scalability, extensible, robust fault-tolerant, minimal maintenance.

IBM Big data platform

- Give a solution which is designed specifically with the needs of the enterprise in the mind.

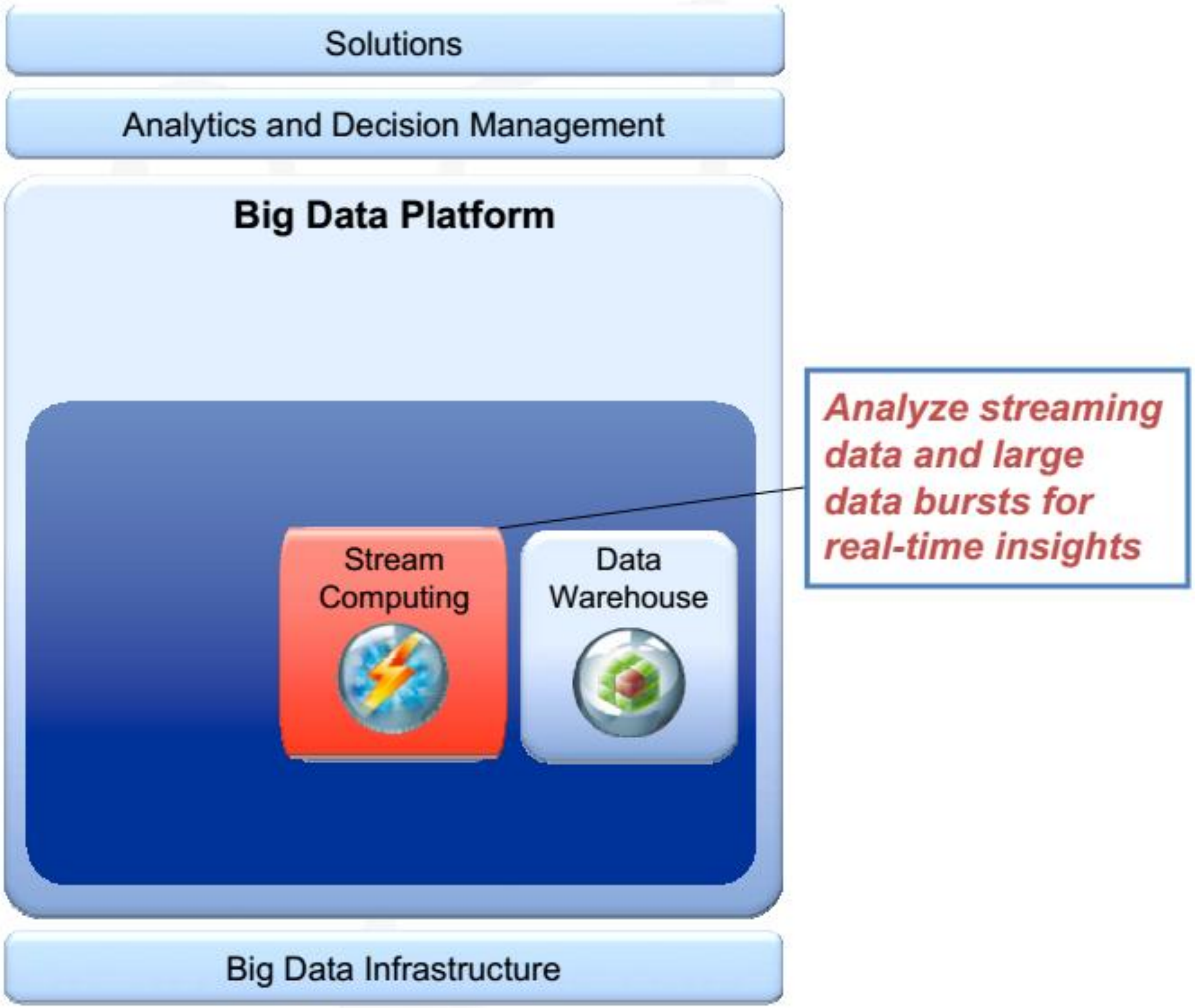


IBM Big data platform



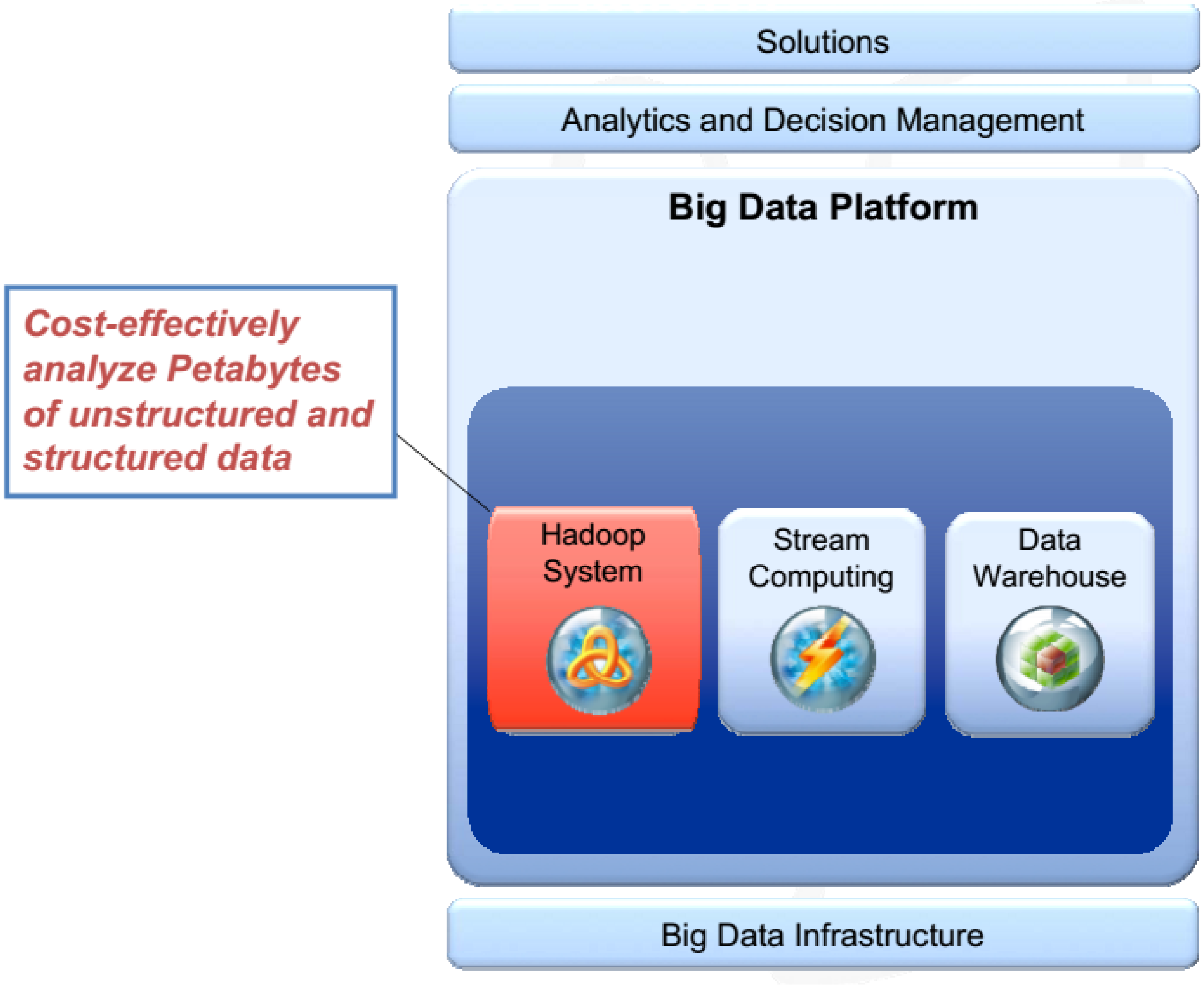
Delivers deep insight with advanced in-database analytics & operational analytics

IBM Big data platform



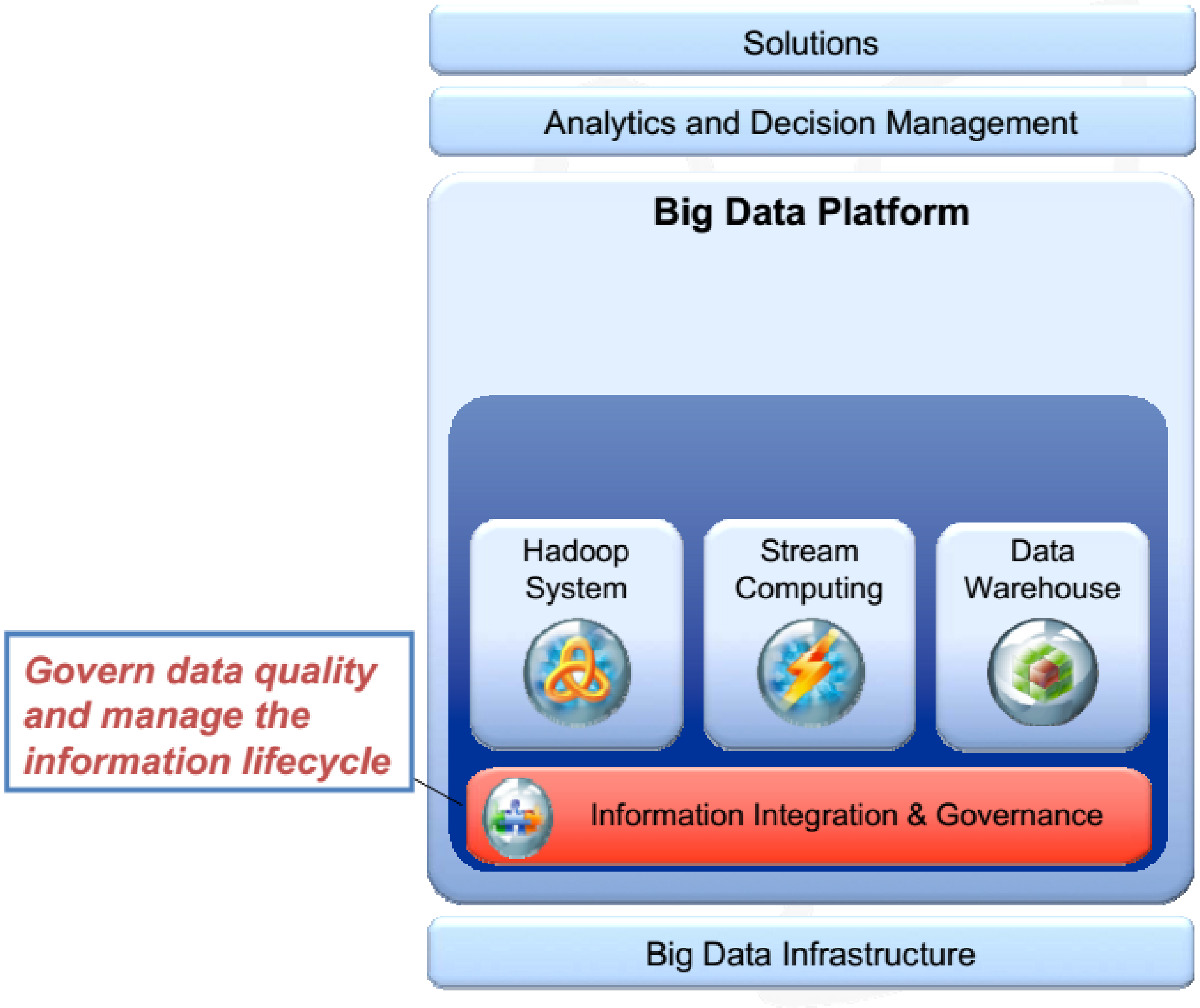
Analyze streaming data and large data bursts for real-time insights

IBM Big data platform



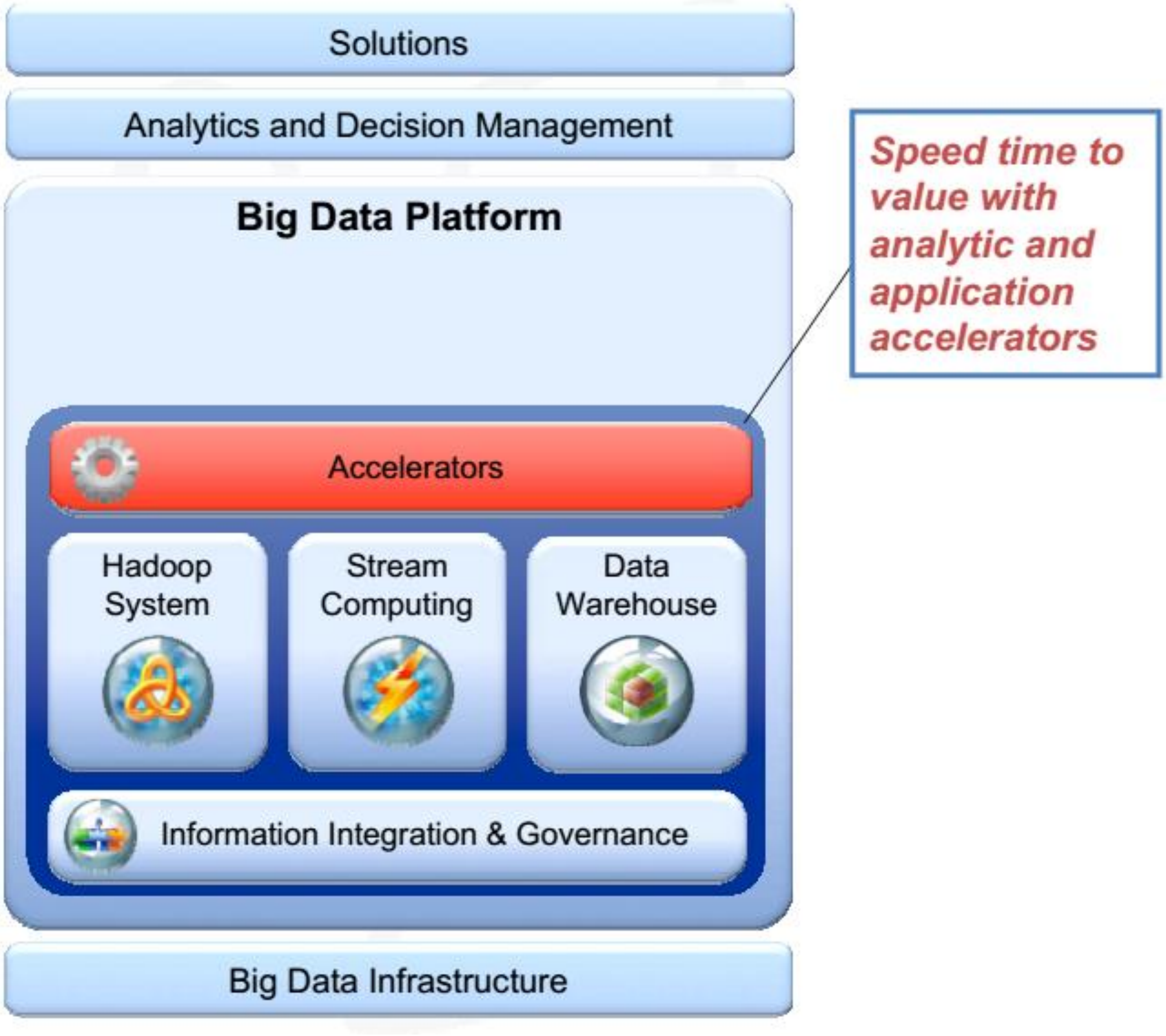
Cost-effectively analyze Petabytes of unstructured and structured data

IBM Big data platform



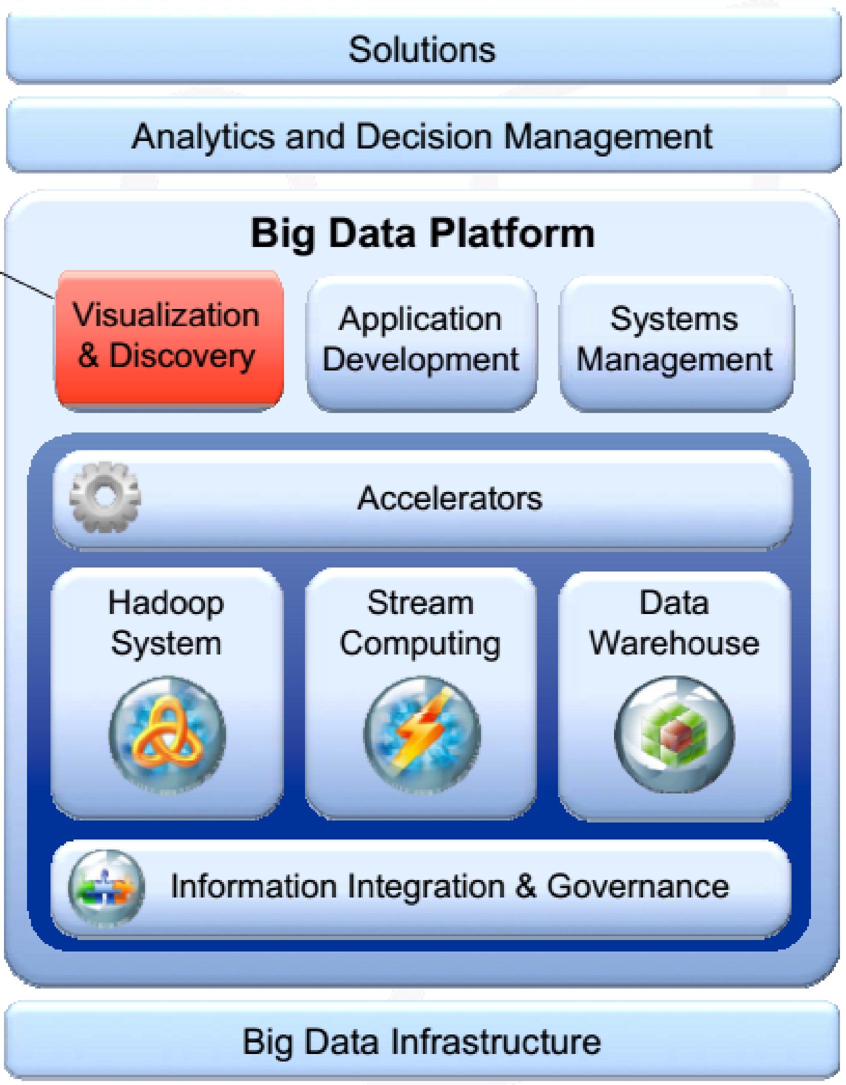
Govern data quality and manage the information lifecycle

IBM Big data platform



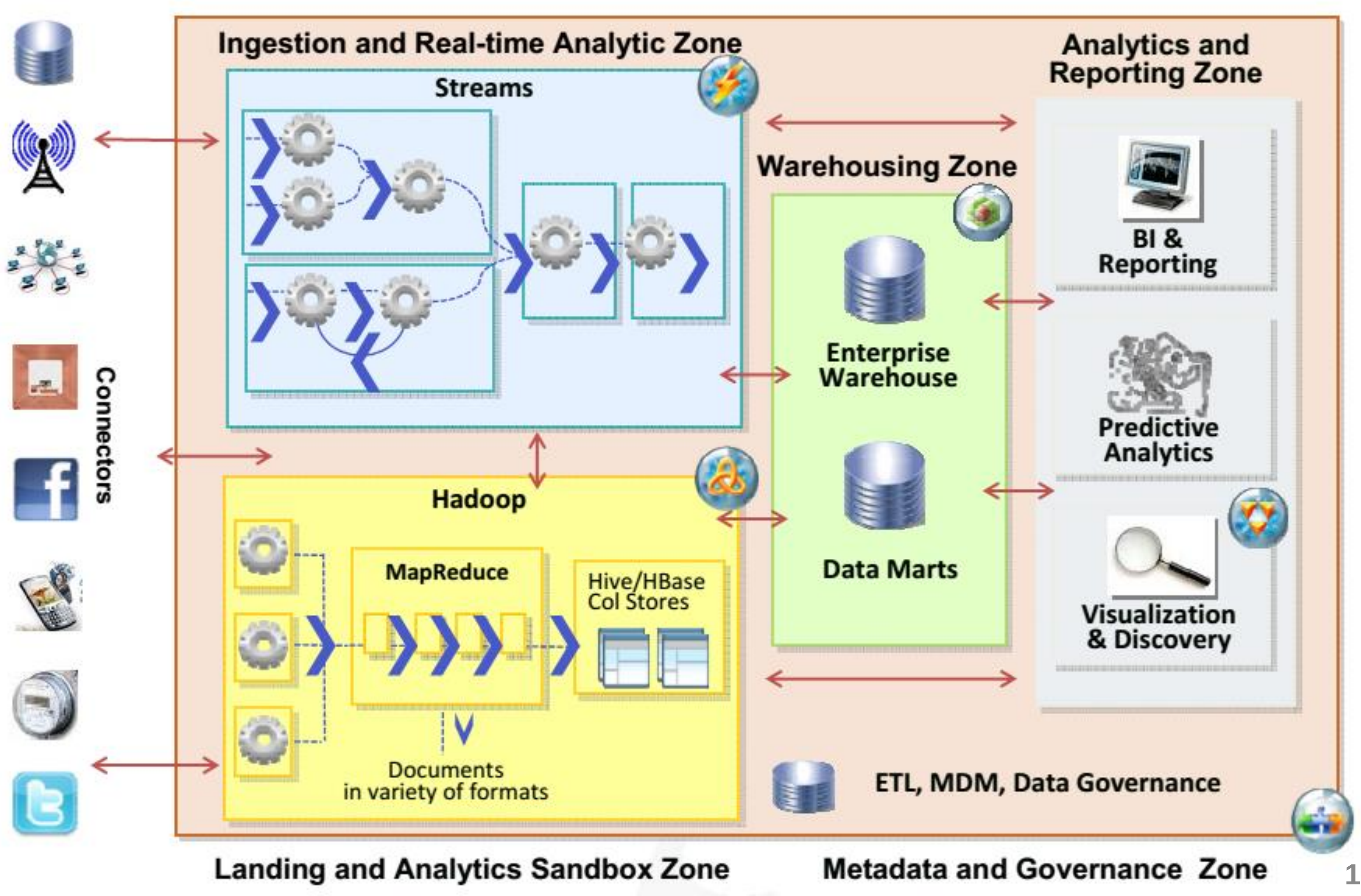
Speed time to value with analytic and application accelerators

IBM Big data platform



Discover, understand, search, and navigate federated sources of big data

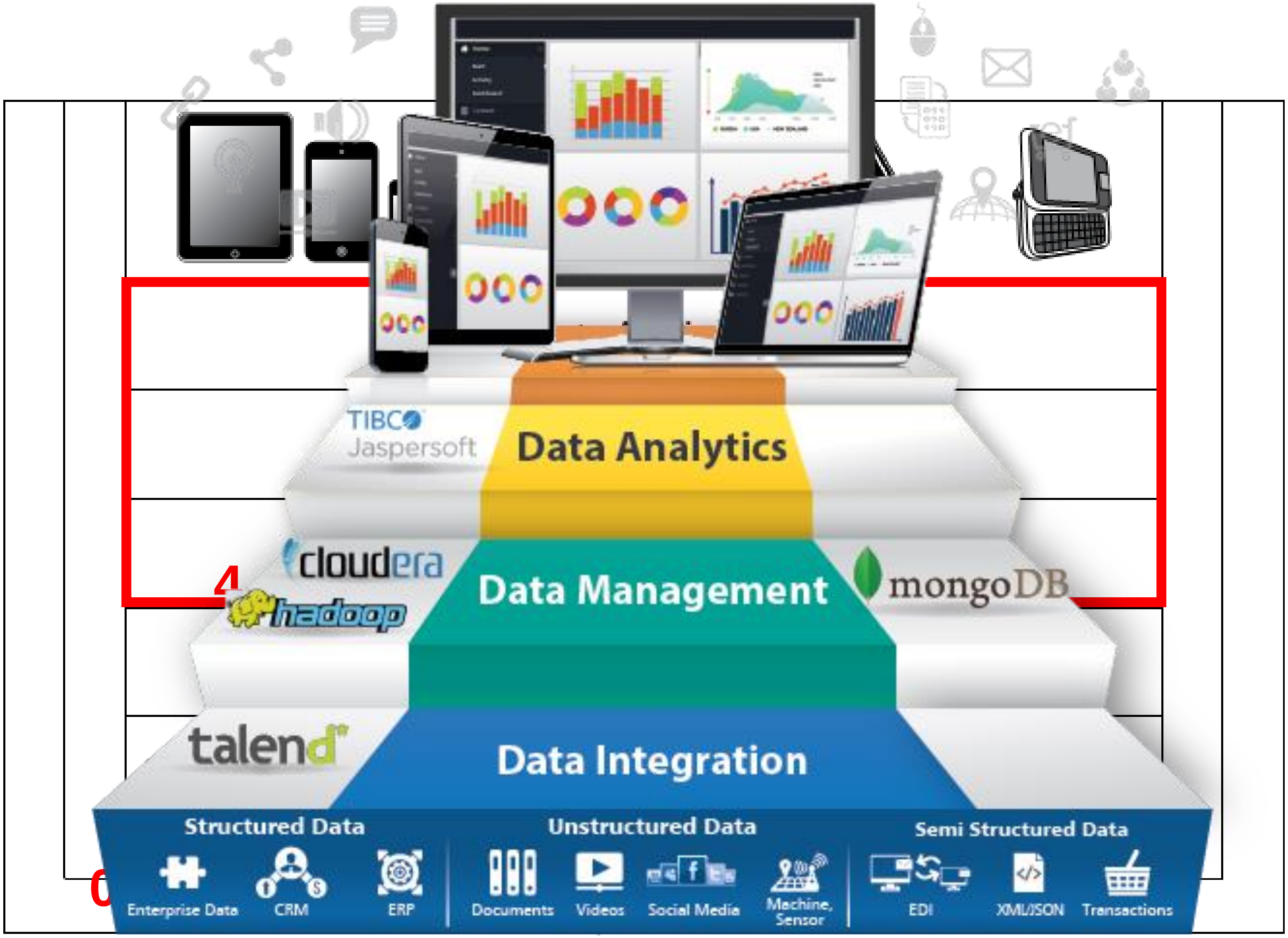
Components in a Big data platform


Ingestion and Real-time Analytic Zone
Streams
Analytics and Reporting Zone
Warehousing Zone
BI & Reporting
Enterprise Warehouse
Predictive Analytics
Connectors
Hadoop
MapReduce
Hive/HBase Col Stores
Data Marts
Visualization & Discovery
Documents in variety of formats
ETL, MDM, Data Governance
Landing and Analytics Sandbox Zone
Metadata and Governance Zone

Digging into Big data technology

Digging deeper, better insights

Big data technology stack

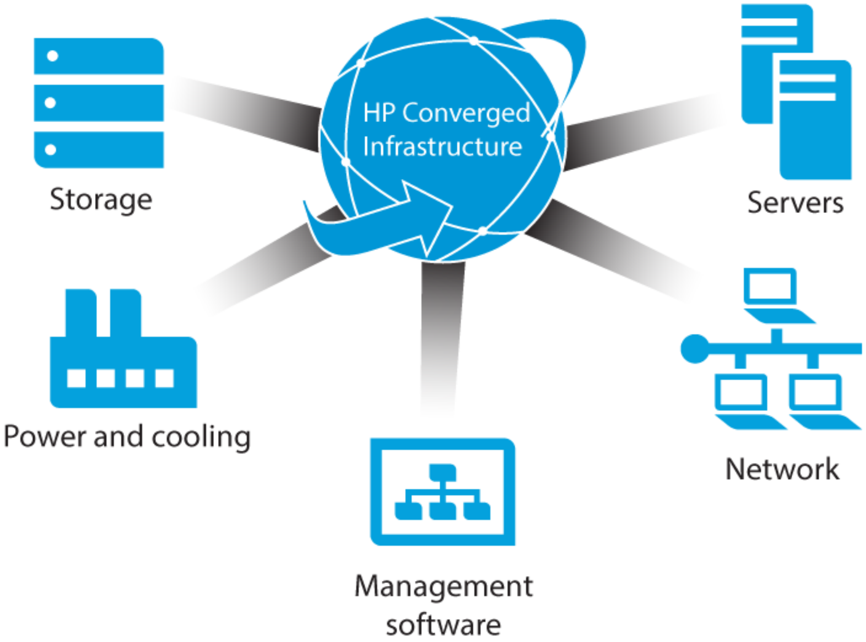


Interfaces and feeds from/to the Internet

Interfaces and feeds from/to internal applications

Layer 0: Redundant physical infrastructure

- The **physical infrastructure** is the lowest level.
  - Hardware, network, etc.



- Your company might already have a data center or made investments in physical infrastructures.
- Hence, you may want to find a way to utilize existing assets.

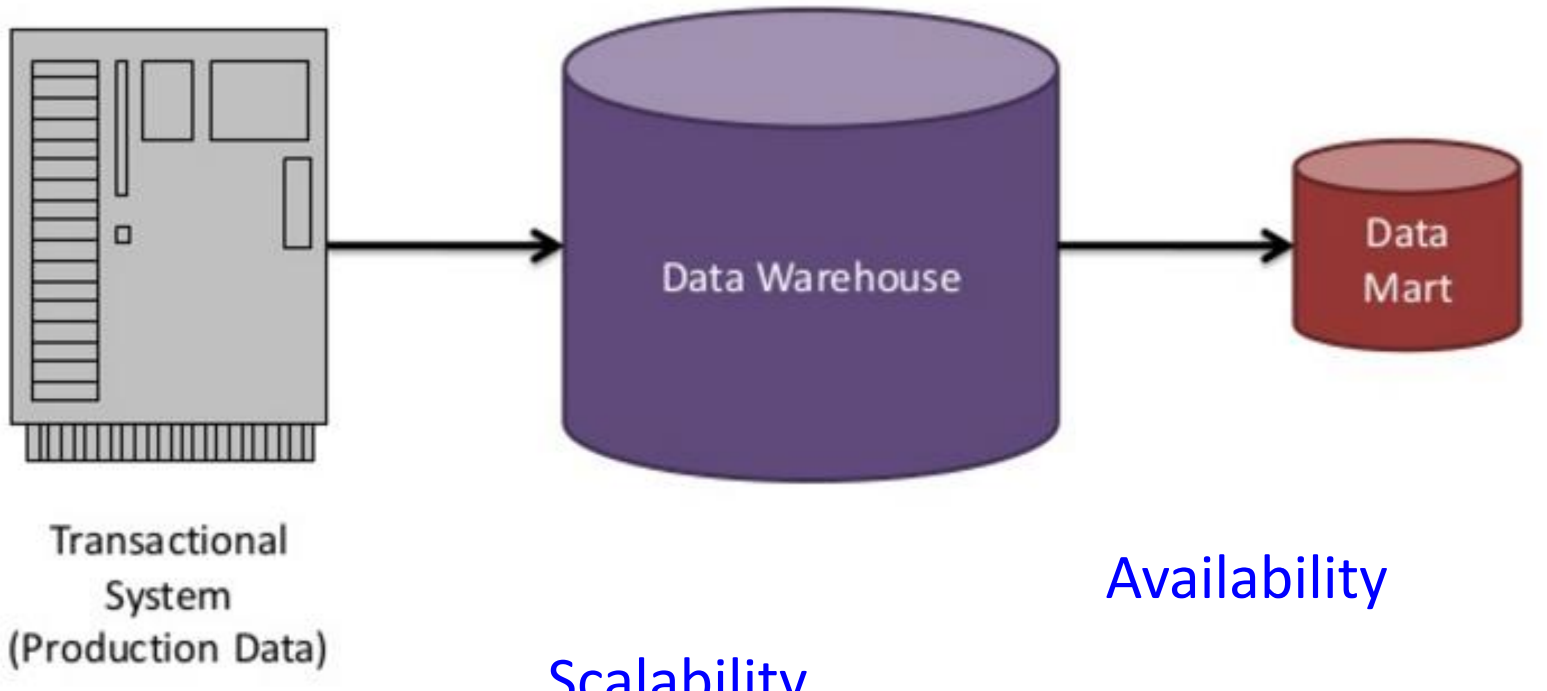
Where most of this began?

- A prioritized list of these principles should include statements about the following

Flexibility

Performance

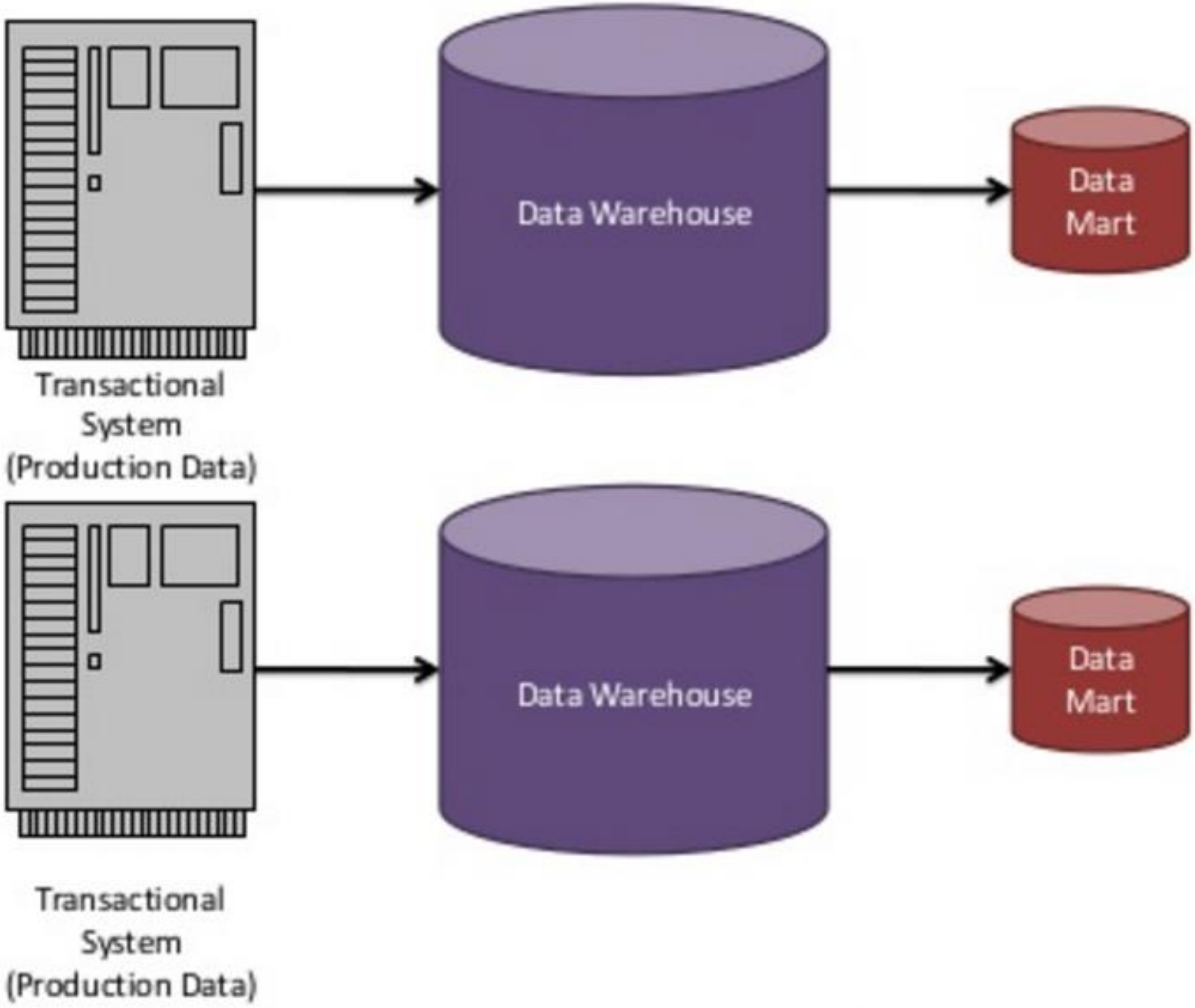
Cost



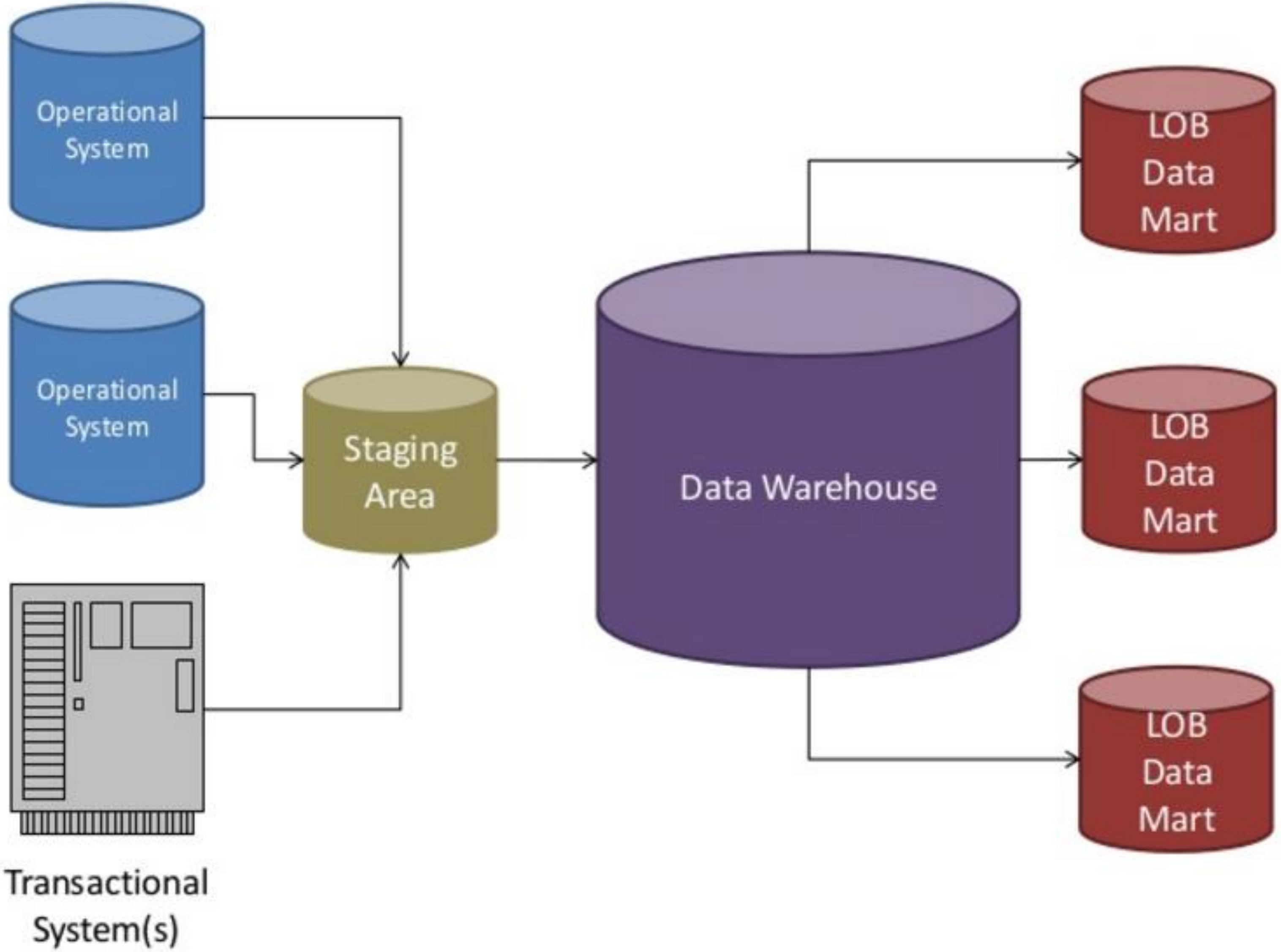
Availability

Scalability

It grows bigger..


Data Warehouse
Data Mart
Transactional System (Production Data)
Data Warehouse
Data Mart
Transactional System (Production Data)

….then very big



Why redundant?

- Most big data implementations need to be highly available.
- That is, networks, servers, and physical storage must be both resilient and redundant.
- A system is resilient to failure or changes when sufficient redundant resources are in place, ready to jump into action.

Layer 1: Security infrastructure

- Security and privacy requirements for big data are similar to those for conventional data environments.
- They must be closely aligned to specific business needs.

Data access

The data should be available only to those who have a legitimate business need for examining or interacting with it.

Application access

Protection from unauthorized usage or access are offered by most APIs.

Data encryption

Most challenging, extremely stress the systems' resources
Encrypt only data elements that require this level of security

Threat detection

The inclusion of mobile devices and social networks exponentially increases both the amount of data and the opportunities for security threats.

Layer 2: Operational databases

- The core of any Big data environment is database engines holding collections of data elements relevant to a business.

Atomicity

If any part of the transaction or the underlying system fails, the entire transaction fails.

Consistency

Only transactions with valid data will be performed.

Isolation

Multiple simultaneous transactions do not interfere with each other. All valid transactions will execute until completed and in the order, they were submitted for processing.

Durability

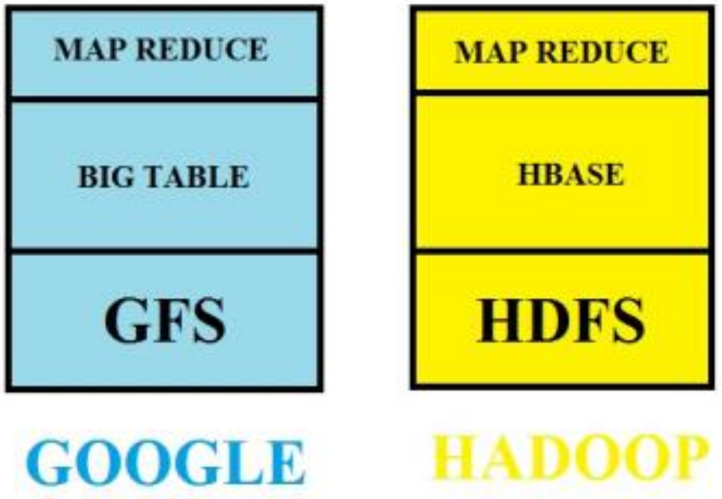
After the data from the transaction is written to the database, it stays there “forever.”

Layer 3: Organizing Data Services and Tools

- Prepare an ecosystem of tools and technologies to gather and assemble data in preparation for further processing
- Technologies in this layer include the following:
  - A distributed file system
  - Serialization services
  - Coordination services
  - Extract, transform, and load (ETL) tools
  - Workflow services

Hadoop, MapReduce and Big Table

- New technologies to store, access, and analyze huge amounts of data



- Proved to be the sparks that led to a new generation of data management.
- Addressing one of the most fundamental problems: the capability of processing massive amounts of data efficiently, cost effectively, and in a timely fashion.

Layer 4: Traditional and advanced analytics

- **What** does your business now **do with all the data** in all its forms to try to make sense of it **for the business**?
  - Managing big data holistically requires different analysis approaches, *depending on the problem being solved*, to help the business to successfully plan.
  - Some analyses will use a traditional data warehouse, while the others will take advantage of advanced predictive analytics.
- **Key techniques:** Analytical data warehouses and data marts, Big data analytics, Reporting and visualization, etc.

Big data platform and analytics software

- Features of Big data platform and analytics software

Data ingestion, Data management, ETL and Warehouse, Hadoop system and Stream Computing

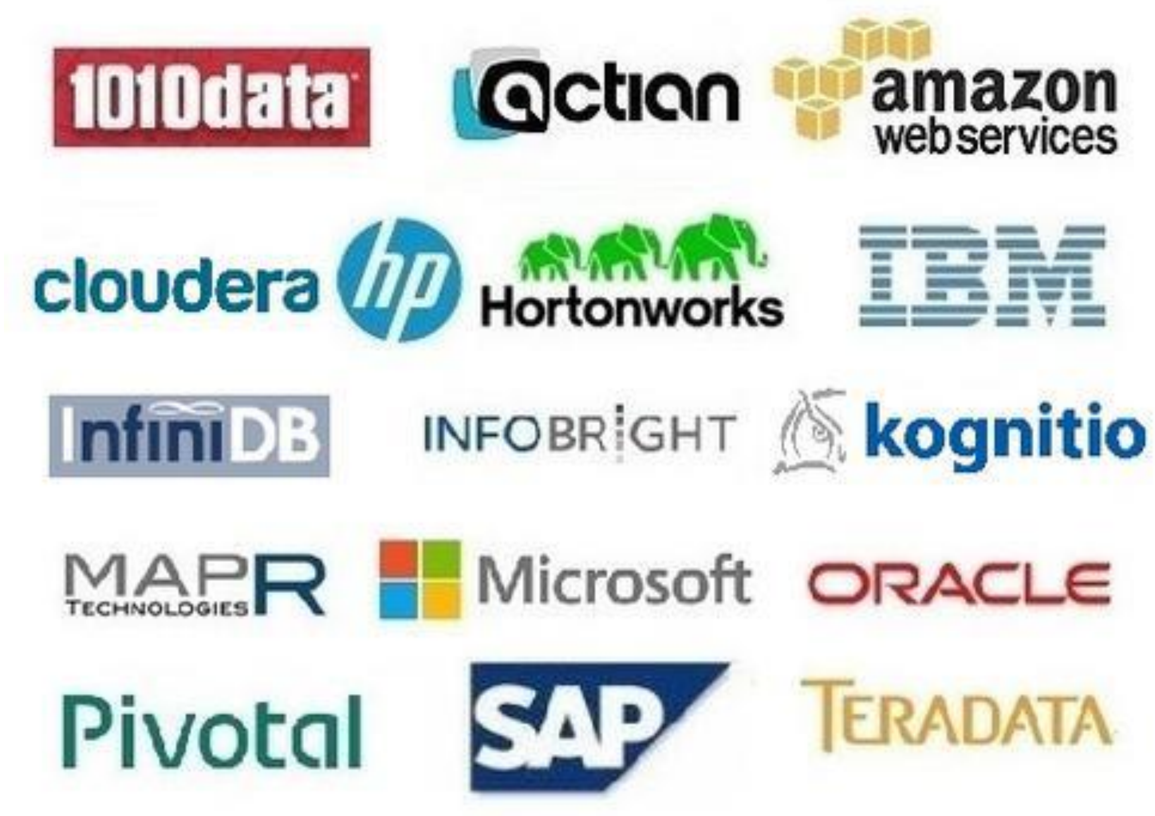
Analytics/Machine learning, Content management, Data integration and governance

Provide efficiency in workplace
Provide accurate data
Give answer to complex questions
It is secure

Source: Pat Research

Big data analytic platform tools

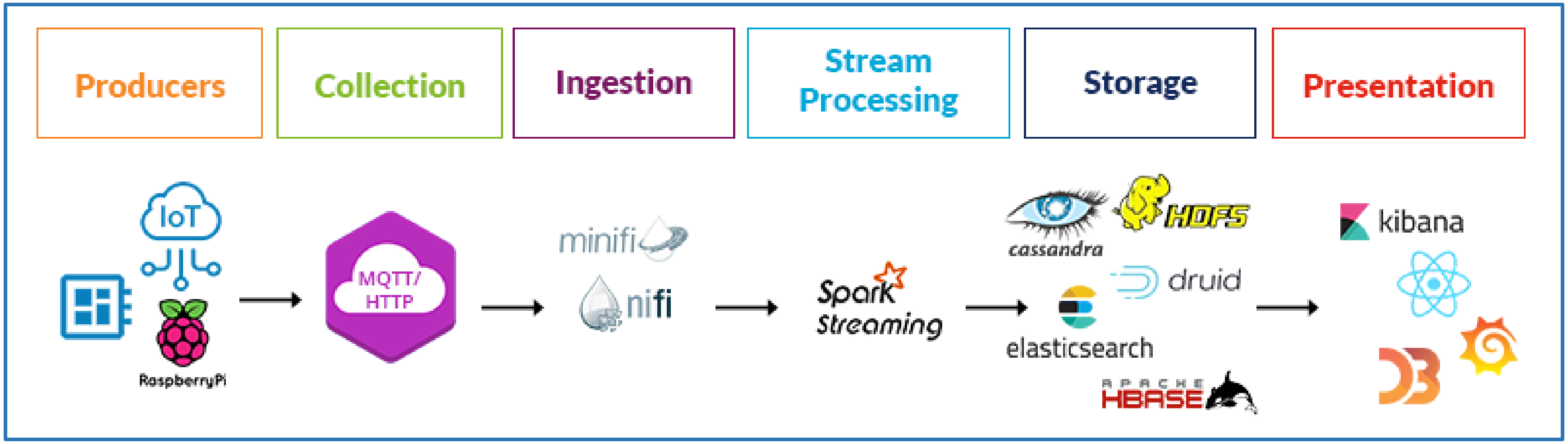
- There are some key Big data analytic platform tools available for enterprise use



XENONSTACK
A Stack Innovator

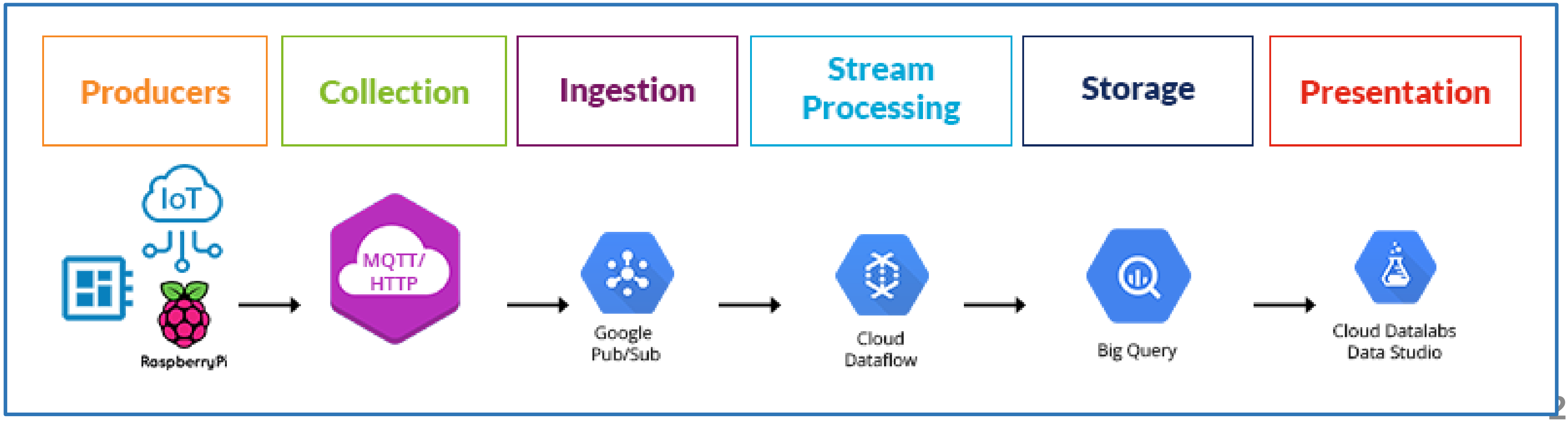
IoT Analytics Platform for Real-Time Data Ingestion

Xenonify IoT Architecture



Producers → Collection → Ingestion → Stream Processing → Storage → Presentation

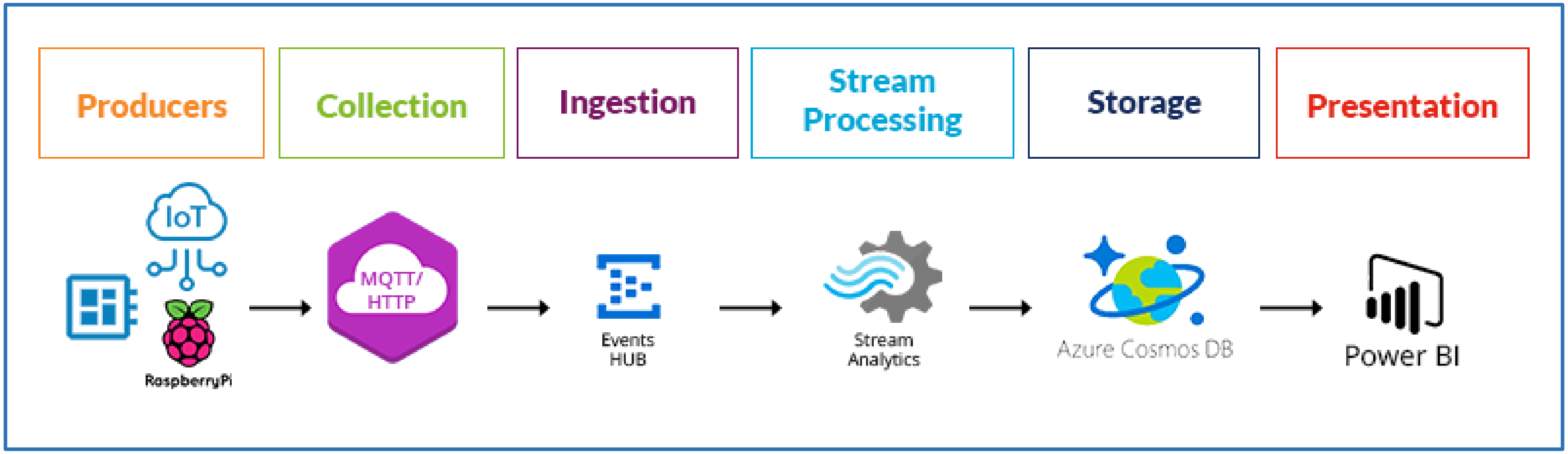
Google IoT Architecture



Producers → Collection → Ingestion → Stream Processing → Storage → Presentation

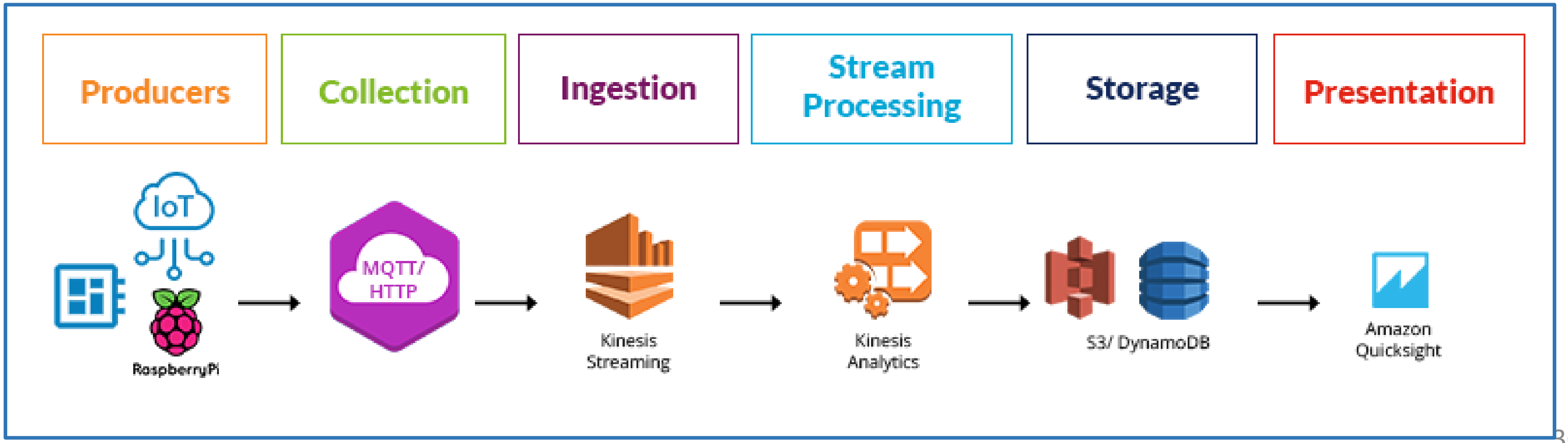
IoT Analytics Platform for Real-Time Data Ingestion

Microsoft Azure IoT Architecture



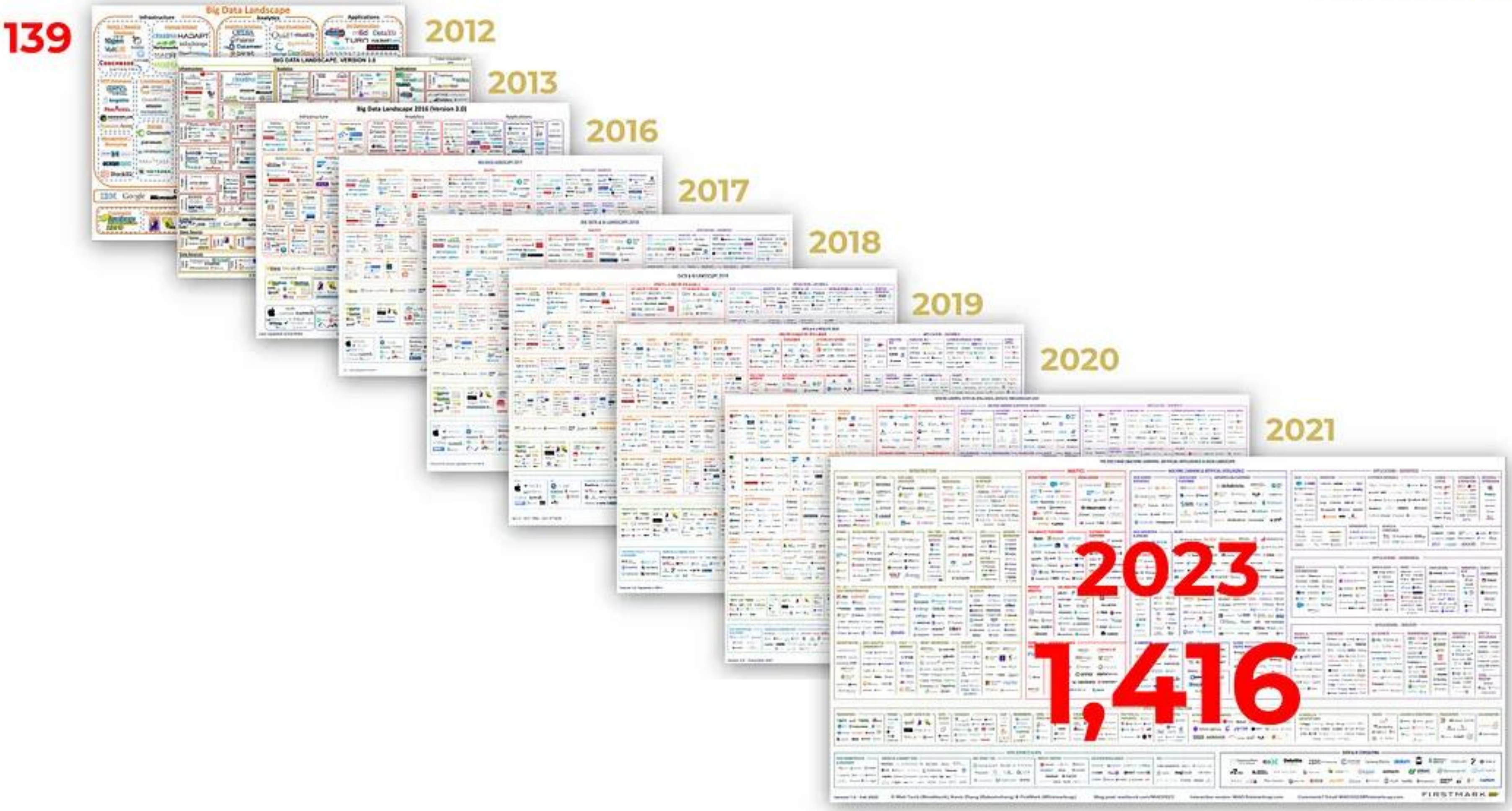
Producers → Collection → Ingestion → Stream Processing → Storage → Presentation

AWS IoT Architecture



Producers → Collection → Ingestion → Stream Processing → Storage → Presentation

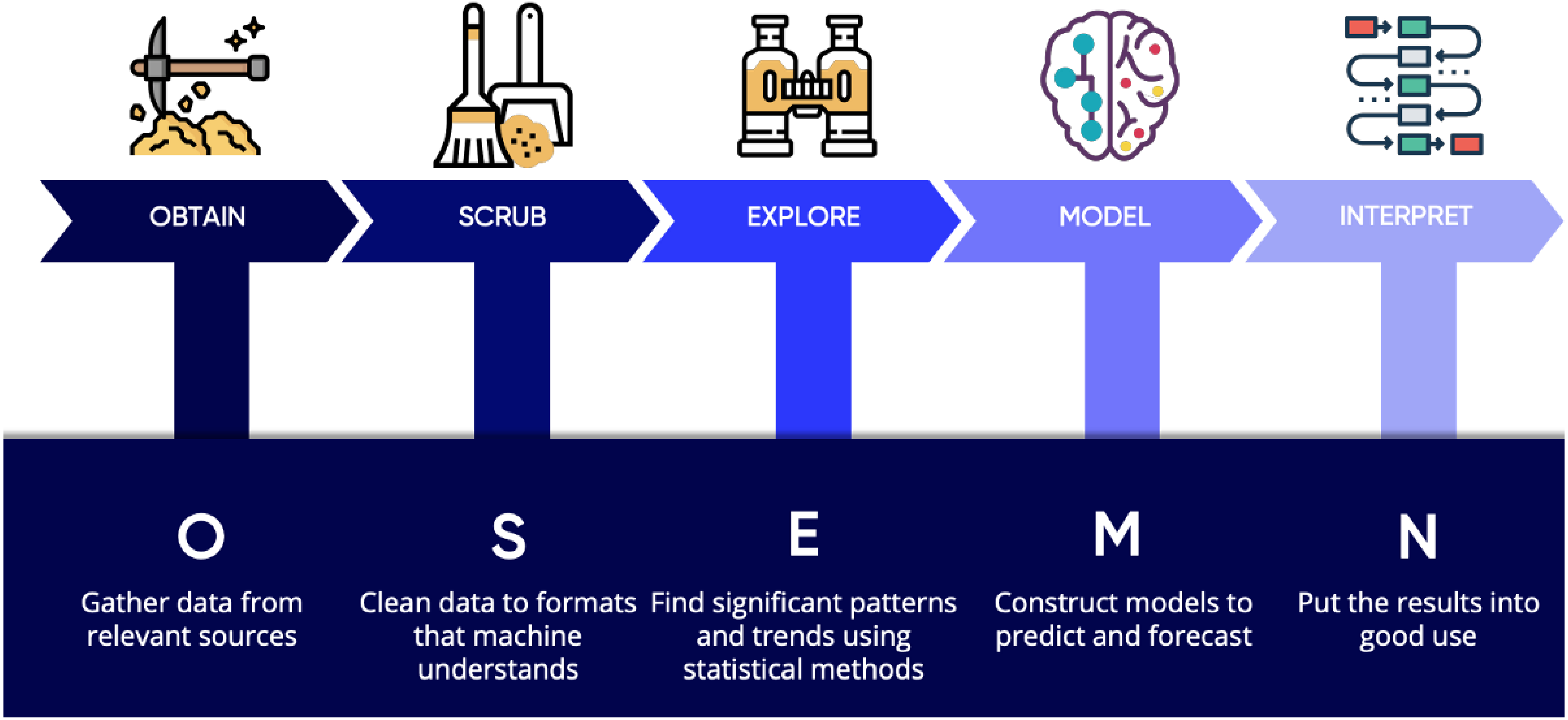
ENGINEERINGANGELS



The 2023 MAD (ML/AI/Data) Landscape

Big data vs. Data science

- In Data science, the data can be of all sizes, which is related to a business or scientific case.
- Big data offers techniques to handle large-scale data at different steps.



OBTAIN

SCRUB

EXPLORE

MODEL

INTERPRET

O

Gather data from relevant sources

S

Clean data to formats that machine understands

E

Find significant patterns and trends using statistical methods

M

Construct models to predict and forecast

N

Put the results into good use

...the end.