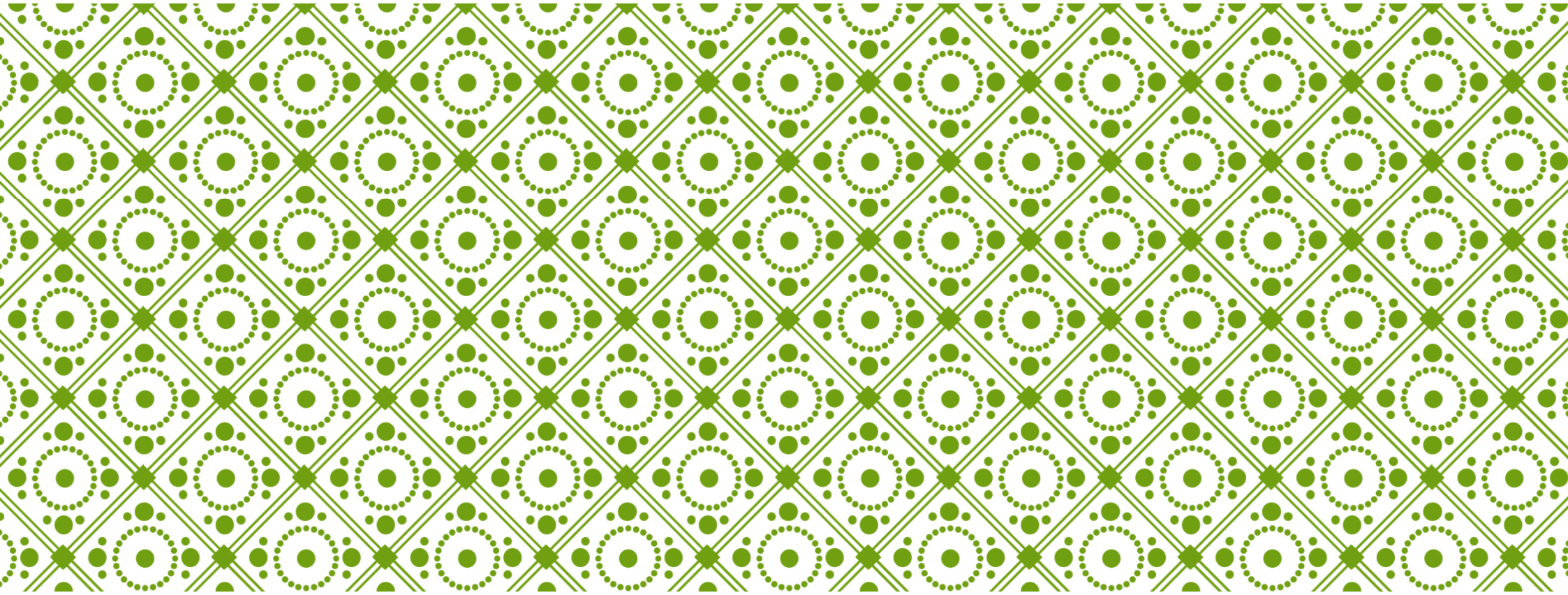


DIGGING INTO BIG DATA TECHNOLOGY COMPONENTS

Indra Aulia
indraaulia@usu.ac.id

RECAP

Big Data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

Characteristics of Big Data: Volume (*How much data?*), Velocity (*How fast that data is processed?*), Variety (*the various data type*), and Veracity (*How accurate is that data in predicting business value?*).

BUILDING A SUCCESSFUL BIG DATA MANAGEMENT ARCHITECTURE

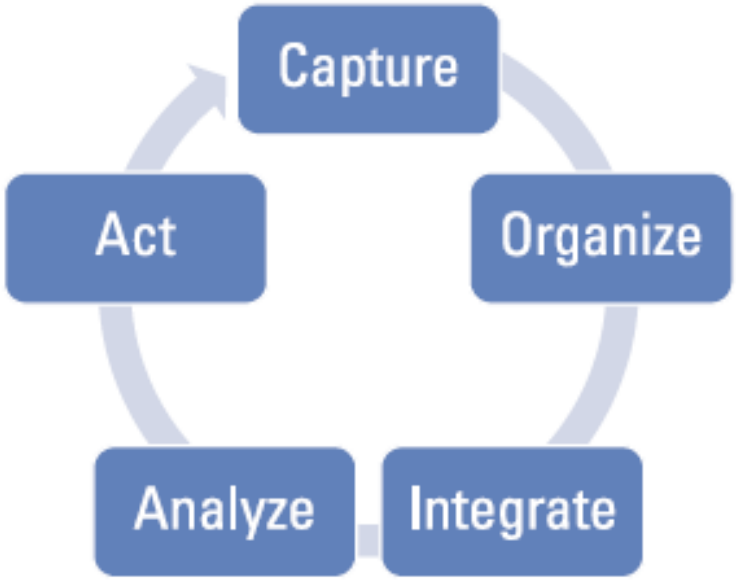


Figure 1-1: The cycle of big data management.

Of course, any foray into big data first needs to start with the problem you're trying to solve. That will dictate the kind of data that you need and what the architecture might look like

BIG DATA ARCHITECTURE

To understand big data, there are question which helps to lay out the components of the architecture.

- How much data will my organization need to manage today and in the future?
- How often will my organization need to manage data in real time or near real time?
- How much risk can my organization afford? Is my industry subject to strict security, compliance, and governance requirements?
- How important is speed to my need to manage data?
- How certain or precise does the data need to be?

BIG DATA ARCHITECTURE

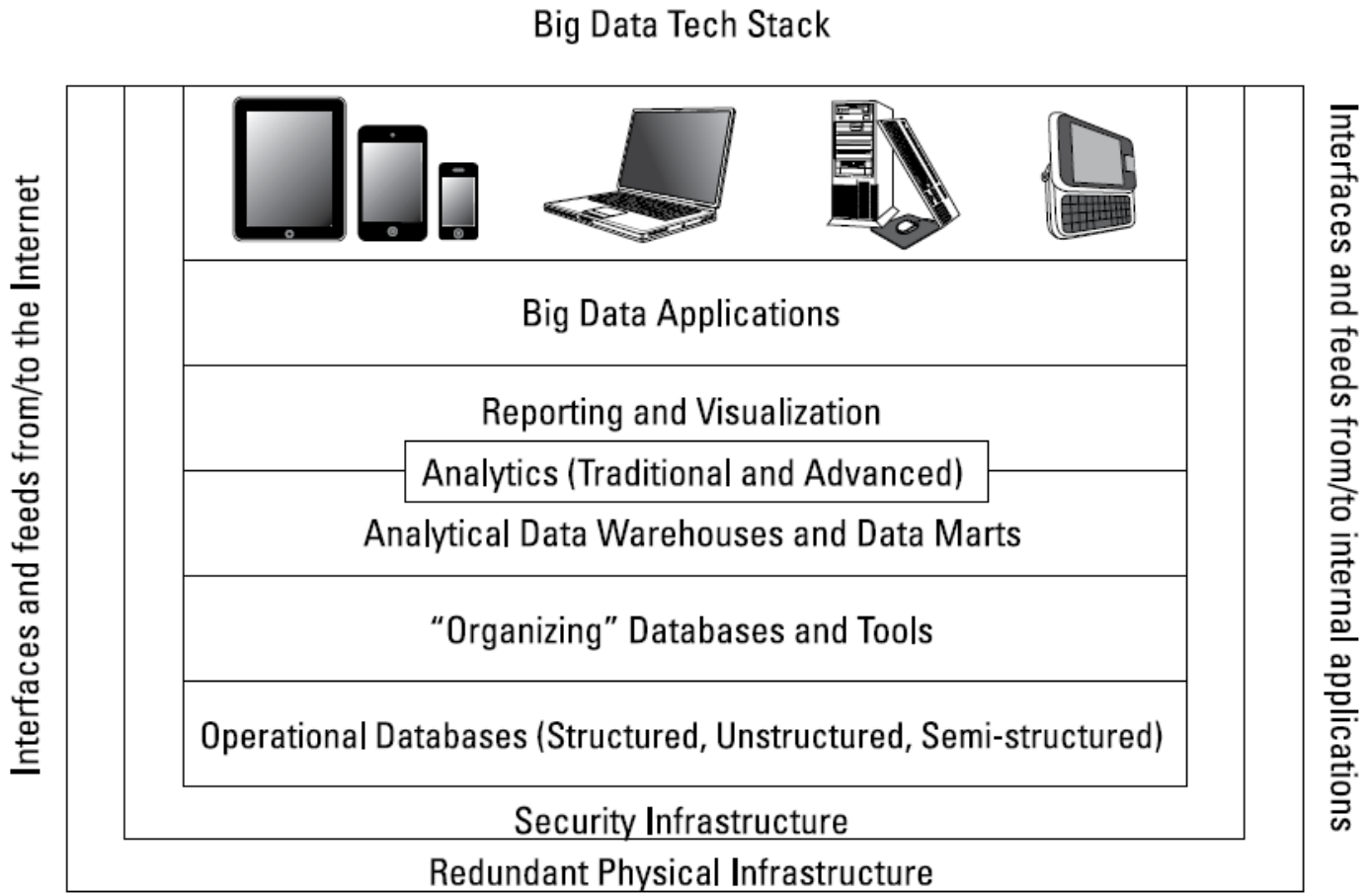


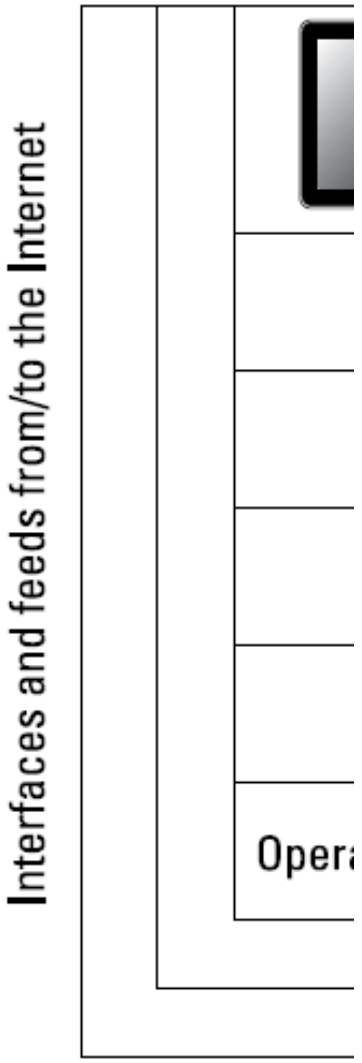
Figure 1-2: The big data architecture.

BIG DATA ARCHITECTURE

INTERFACE AND FEEDS

In fact, what makes big data big is the fact that it relies on picking up lots of data from lots of sources.

Open application programming interfaces (APIs) will be core to any big data architecture



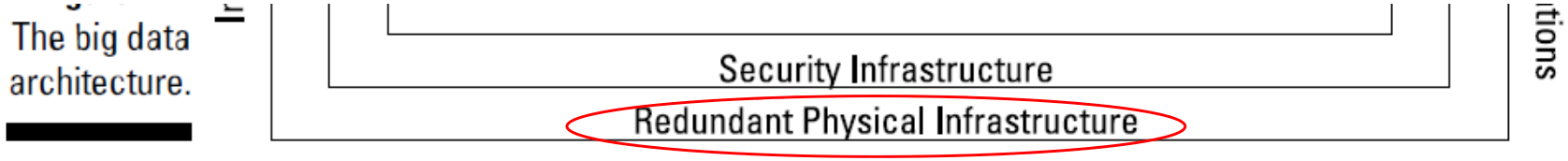
BIG DATA ARCHITECTURE
PHYSICAL INFRASTRUCTURE

The supporting physical infrastructure is fundamental to the operation and scalability of a big data architecture.

At the lowest level of the stack is the physical infrastructure — the hardware, network, and so on.

To support an unanticipated or unpredictable volume of data, a physical infrastructure for big data has to be different than that for traditional data.

The physical infrastructure is based on **a distributed computing model.**



The big data architecture.

BIG DATA ARCHITECTURE
PHYSICAL INFRASTRUCTURE

As you start to think about your big data implementation, it is important to have some overarching principles that you can apply to the approach. A prioritized list of these principles should include statements about the following:

Performance: How responsive do you need the system to be?

Availability: Do you need a 100 percent uptime guarantee of service? How long can your business wait in the case of a service interruption or failure?

Scalability: How big does your infrastructure need to be? How much disk space is needed today and in the future? How much computing power do you need?

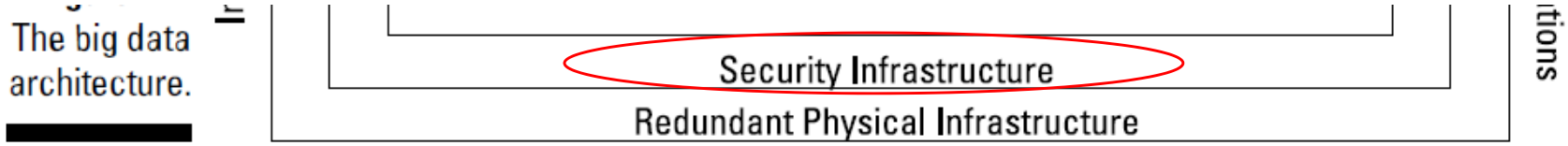
Flexibility: How quickly can you add more resources to the infrastructure? How quickly can your infrastructure recover from failures?

Cost: What can you afford?

BIG DATA ARCHITECTURE
SECURITY INFRASTRUCTURE

Security and privacy requirements for big data are similar to the requirements for conventional data environments.

The security requirements have to be closely aligned to specific business needs.



The big data architecture.

BIG DATA ARCHITECTURE
SECURITY INFRASTRUCTURE

Some unique challenges arise when big data becomes part of the strategy, which we briefly describe in this list:

Data access: User access to raw or computed big data has about the same level of technical requirements as non-big data implementations.

Application access: Application access to data is also relatively straightforward from a technical perspective.

Data encryption: Data encryption is the most challenging aspect of security in a big data environment.

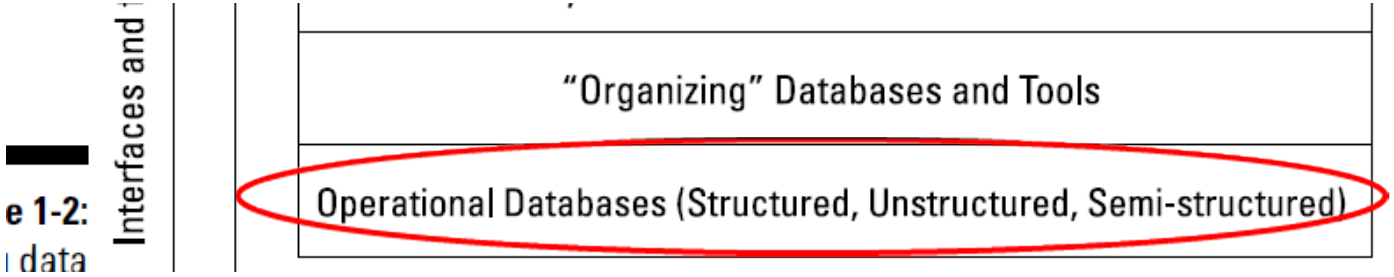
Threat detection: The inclusion of mobile devices and social networks exponentially increases both the amount of data and the opportunities for security threats.

BIG DATA ARCHITECTURE

OPERATIONAL DATABASES

At the core of any big data environment are the database engines containing the collections of data elements relevant to your business.

These engines need to be fast, scalable, and rock solid.



BIG DATA ARCHITECTURE
OPERATIONAL DATABASES

Database designers describe this behavior with the acronym ACID. It stands for:

Atomicity: A transaction is "all or nothing" when it is atomic. If any part of the transaction or the underlying system fails, the entire transaction fails.

Consistency: Only transactions with valid data will be performed on the database. If the data is corrupt or improper, the transaction will not complete and the data will not be written to the database.

Isolation: Multiple, simultaneous transactions will 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."

BIG DATA ARCHITECTURE

OPERATIONAL DATABASES

Table 4-1 Important Characteristics of SQL and NoSQL Databases

Engine	*Query Language*	*MapReduce*	*Data Types*	*Transactions*	*Examples*
Relational	SQL, Python, C	No	Typed	ACID	PostgreSQL, Oracle, DB/2
Columnar	Ruby	Hadoop	Predefined and typed	Yes, if enabled	HBase
Graph	Walking, Search, Cypher	No	Untyped	ACID	Neo4J
Document	Commands	JavaScript	Typed	No	MongoDB, CouchDB
Key-value	Lucene, Commands	JavaScript	BLOB, semityped	No	Riak, Redis

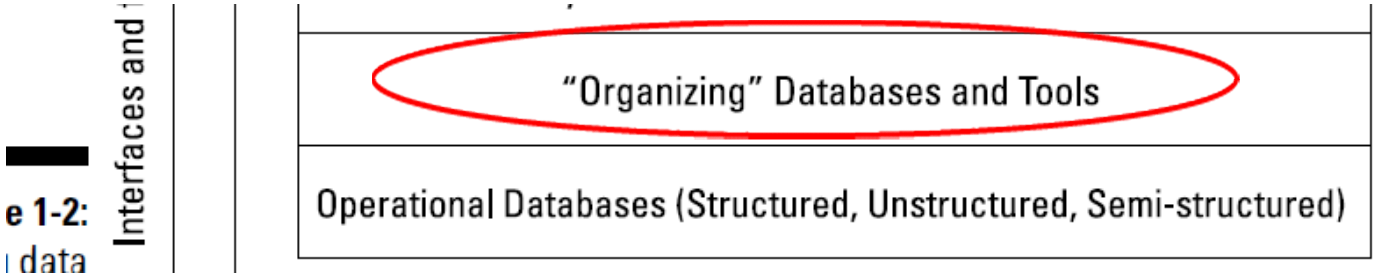
BIG DATA ARCHITECTURE

ORGANIZING DATA SERVICES AND TOOLS

Organizing data services and tools capture, validate, and assemble various big data elements into contextually relevant collections.

Because big data is massive, techniques have evolved to process the data efficiently and seamlessly.

MapReduce is one heavily used technique.



BIG DATA ARCHITECTURE
ORGANIZING DATA SERVICES AND TOOLS

A distributed file system: Necessary to accommodate the decomposition of data streams and to provide scale and storage capacity

Serialization services: Necessary for persistent data storage and multilanguage remote procedure calls (RPCs)

Coordination services: Necessary for building distributed applications (locking and so on)

Extract, transform, and load (ETL) tools: Necessary for the loading and conversion of structured and unstructured data into Hadoop

Workflow services: Necessary for scheduling jobs and providing a structure for synchronizing process elements across layers

BIG DATA ARCHITECTURE

ANALYTICAL DATA WAREHOUSES

The data warehouse , and its companion the data mart, have long been the primary techniques that organizations use to optimize data to help decision makers.

Typically, data warehouses and marts contain normalized data gathered from a variety of sources and assembled to facilitate analysis of the business.

BIG DATA ARCHITECTURE
ANALYTICAL DATA WAREHOUSES

Because many data warehouses and data marts are comprised of data gathered from various sources within a company, the costs associated with the cleansing and normalizing of the data must also be addressed.

With big data, you find some key differences:

- Traditional data streams (from transactions, applications, and so on) can produce a lot of disparate data.
- Dozens of new data sources also exist, each of them needing some degree of manipulation before it can be timely and useful to the business.
- Content sources will also need to be cleansed, and these may require different techniques than you might use with structured data.

BIG DATA ARCHITECTURE

BIG DATA ANALYTICS

Existing analytics tools and techniques will be very helpful in making sense of big data. However, there is a catch.

The algorithms that are part of these tools have to be able to work with large amounts of potentially real-time and disparate data.

BIG DATA ARCHITECTURE
BIG DATA ANALYTICS

Three classes of tools in this layer of Big Data Analytics:

Reporting and dashboards: These tools provide a “user-friendly” representation of the information from various sources.

Visualization: These tools are the next step in the evolution of reporting.

Analytics and advanced analytics: These tools reach into the data warehouse and process the data for human consumption.

BIG DATA ARCHITECTURE

BIG DATA APPLICATIONS

Custom and third-party applications offer an alternative method of sharing and examining big data sources.

The most prevalent categories as of this writing are log data applications (Splunk, Loggly), ad/media applications (Bluefin, DataXu), and marketing applications (Bloomreach, Myrrix).