



- 1. Overview of BIG data**
- 2. Overview of R**
- 3. Introduction to R**
- 4. Data Structures**
- 5. Programming Concepts**
- 6. Graphs**



1.1 Introduction to BIG data






1.1 Introduction to BIG data – continued

- In Feb 2001, Doug Laney, an analyst with the Meta Group, publishes a research note titled “3D Data Management: Controlling Data volume, Velocity, and Variety.”
- Big data is the capability to manage a huge volume of disparate data, at the right speed, and within the right time frame to allow real-time analysis and reaction.



1.1 Introduction to BIG data - continued

- Big data enables organizations to store, manage, and manipulate vast amounts of data at the right speed and at the right time to gain the right insights.
-  eBay.com uses two data warehouses at 7.5 petabytes and 40PB as well as a 40PB Hadoop cluster for research, customer recommendations, and merchandising.

<http://www.itnews.com.au/News/342615,inside-ebay8217s-90pb-data-warehouse.aspx>



1.2 Defining BIG data

In 2012, Gartner defined BIG data as follows:

Big data is

- ✓ high volume,
- ✓ high velocity, and/or
- ✓ high variety information assets

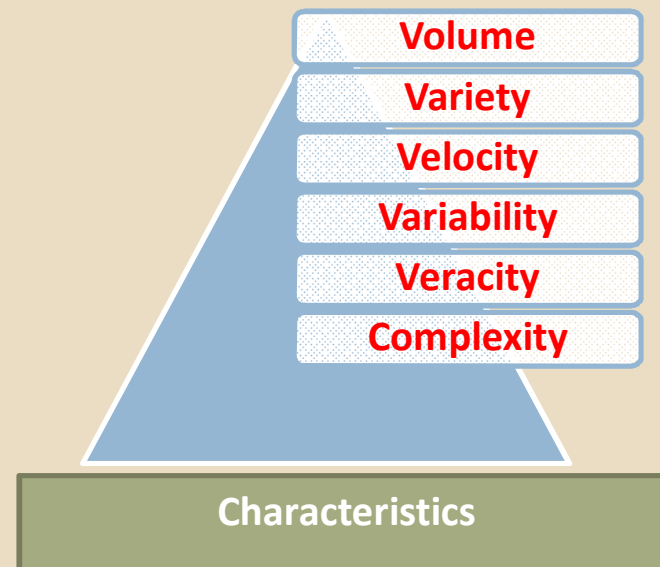
that require new forms of processing to

- ✓ enable decision making,
- ✓ insight discovery and
- ✓ process optimization



1.3 Characteristics of BIG data

1. **Volume** is about the quantity of data
2. **Variety** is about the category to which the data belongs to
3. **Velocity** is about the speed of data generation
4. **Variability** is about the inconsistency shown by the data
5. **Veracity** is about the quality of data
6. **Complexity** arises especially for large volumes from multiple sources



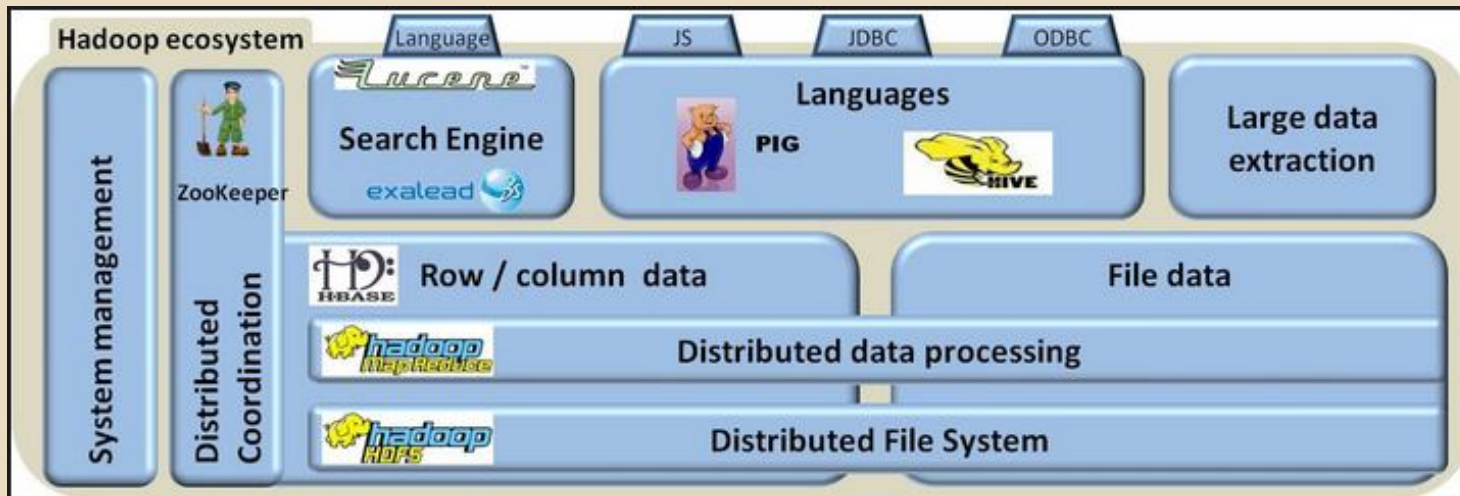


1.4 Why BIG data?

1. Better management of data
2. Benefit from speed, capacity and scalability of cloud storage
3. Improvement in your data analysis methods
4. Explore new business opportunities
5. End users can visualize data



1.5 Tools used in BIG data

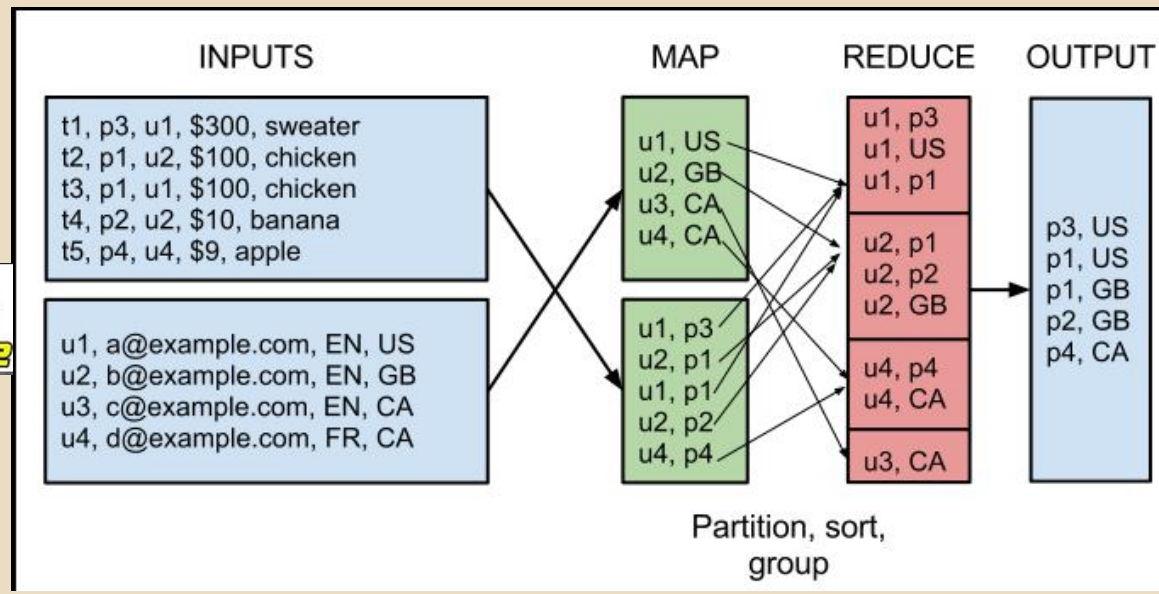


1. Hadoop HDFS – The Hadoop Distributed File System (HDFS) is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications.



1.5 Tools used in BIG data - continued

2. Hadoop mapReduce was designed by Google as a way of efficiently executing a set of functions against a large amount of data in batch mode.





1.5 Tools used in BIG data - continued

2. MapReduce is designed for managing and processing big data; three major features:

- 1) Single and easy programming model*
- 2) automatic and linear scalability and*
- 3) built-in fault tolerance*



1.5 Tools used in BIG data - continued

3. HBASE: is a distributed, non-relational (columnar) database that utilizes HDFS as its persistence store.



Column Family: User

rowid	Col_name	ts	Col_value
u1	name	v1	Ricky
u1	email	v1	ricky@gmail.com
u1	email	v2	ricky@yahoo.com
u2	name	v1	Sam
u2	phone	v1	650-3456

Column Family: Social

rowid	Col_name	ts	Col_value
u1	friend	v1	u10
u1	friend	v1	u13
u2	friend	v1	u10
u2	classmate	v1	u15

- One File per Column Family
- Data inside file is physically sorted
- Sparse: NULL cell does not materialize



1.5 Tools used in BIG data - continued

4a. PIG is an interactive, or script-based, execution environment supporting PIG Latin, a language used to express data flows.

4b. HIVE is a batch-oriented, batch-warehousing layer built on the core elements of Hadoop (HDFS and MapReduce).

It provides users who know SQL with a simple SQL-like implementation called HiveQL, without sacrificing access via mappers and reducers.



1.5 Tools used in BIG data - continued



5. Search engine: Exalead's product, CloudView Solution is an infrastructure-level search and information access platform used for both online and enterprise Search-Based Applications (SBA) as well as enterprise search.

6. ZooKeeper: is a Hadoop's way of coordinating all the elements of the distributed applications.





1.6 Applications of BIG data

➤ Healthcare

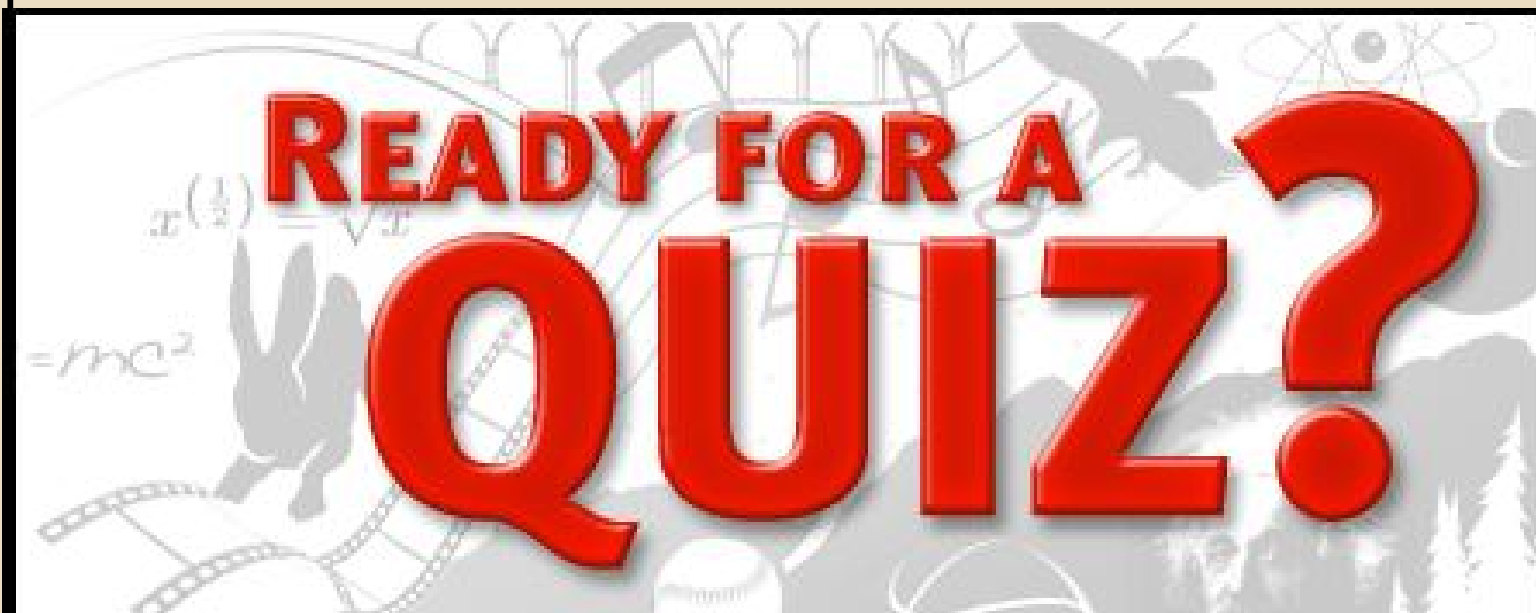
- a. *Evidence based Healthcare Models driven by "Health outcomes"*
- b. *Provider & Staff shortage demands workforce productivity & Efficiency*
- c. *Patient centered Care*
- d. *Disease Management*



➤ Retail

- a. *Behavior profiling*
- b. *Preference modeling*
- c. *Affinity tracking*
- d. *Brand Switching*
- e. *Customer Loyalty*







1) _____ is designed for managing and processing big data; three major features:

- ☐ Single and easy programming model
- ☐ automatic and linear scalability and
- ☐ built-in fault tolerance

The missing word is:

- a) HDFS**
- b) HBase**
- c) Zookeeper**
- d) MapReduce**

2) HBASE is a distributed, _____ database that utilizes HDFS as its persistence store.

The missing word is:

- a) non-relational (columnar)**
- b) relational (row-based)**
- c) All of the above**
- d) None of the above**



3) MapReduce was devised by _____ .

The missing word is:

- a) Apple*
- b) Google*
- c) Microsoft*
- d) None of the above*

4) Batch processing is used when

- i) response time should be short
- ii) data processing is to be carried out at periodic intervals
- iii) transactions are in batches
- iv) transactions do not occur periodically

- a) i, ii*
- b) i, iii, iv*
- c) ii, iii*
- d) i, ii, iii*



ANSWERS

- 1) **d) MapReduce** is designed for managing and processing big data; three major features:
 - ☐ Single and easy programming model
 - ☐ automatic and linear scalability and
 - ☐ built-in fault tolerance
- 2) HBASE is a distributed, **a) non-relational (columnar)** database that utilizes HDFS as its persistence store.
- 3) MapReduce was devised by **b) Google**
Batch processing is used when
 - i) response time should be short
 - ii) data processing is to be carried out at periodic intervals
 - iii) transactions are in batches
 - iv) transactions do not occur periodically

c) ii, iii