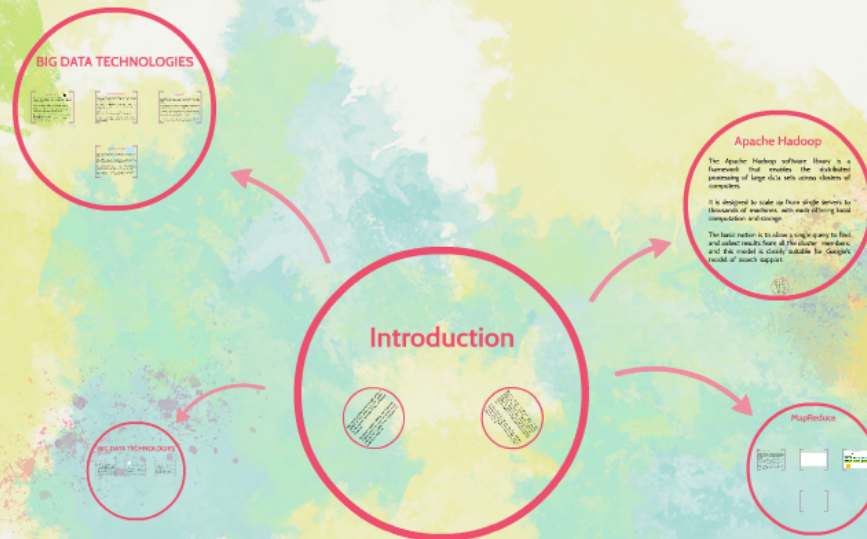
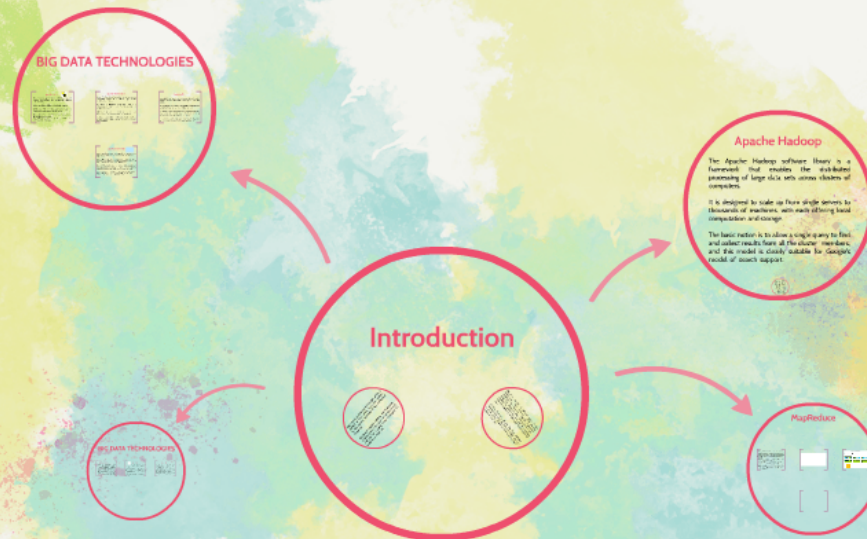


BIG DATA ANALYTICS Harvard Case Solution & Analysis



BIG DATA ANALYTICS

BIG DATA ANALYTICS Harvard Case Solution & Analysis



BIG DATA ANALYTICS

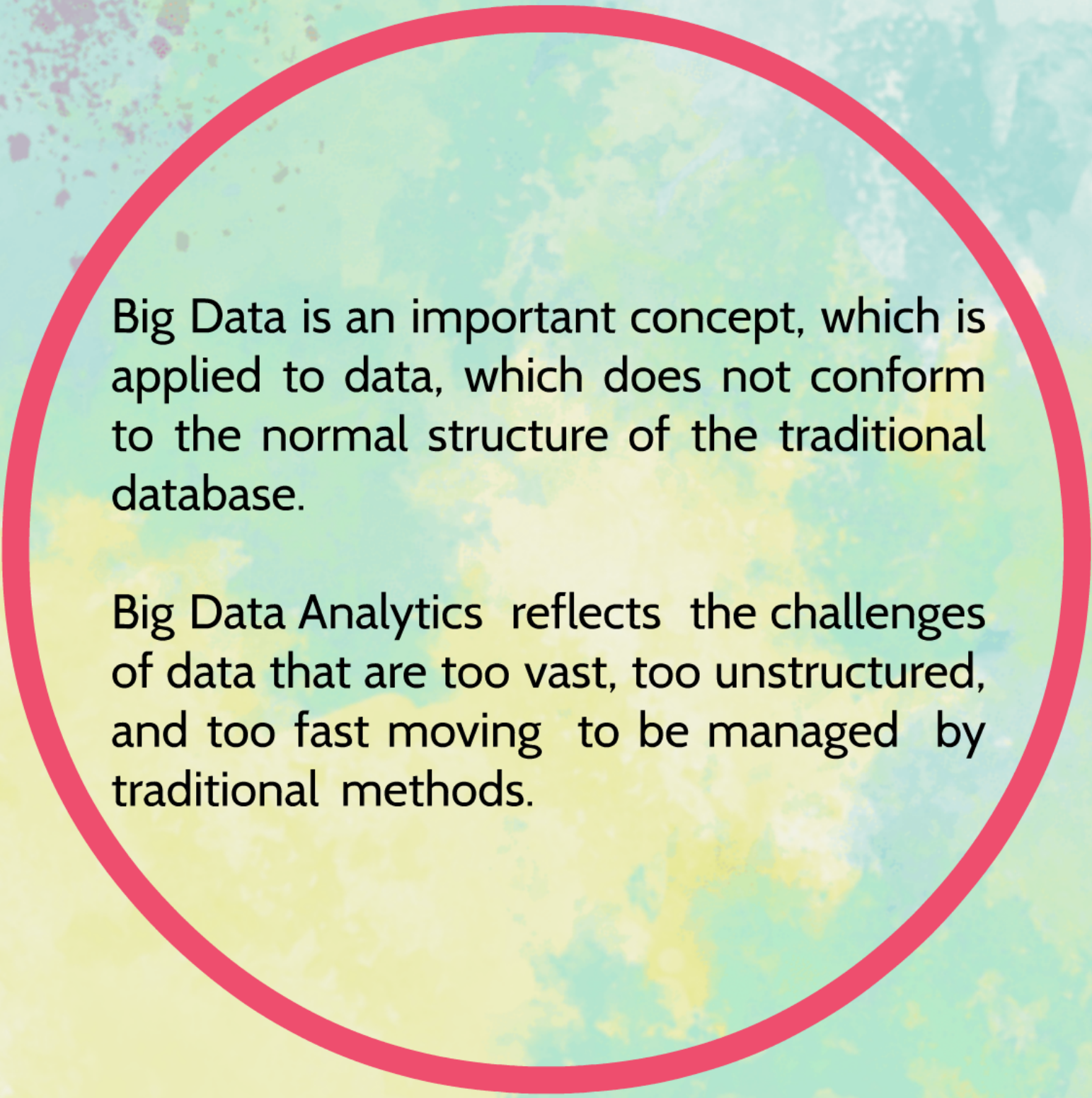
Introduction

Big Data is an important concept, which is applied to data, which does not conform to the normal structure of the traditional database.

Big Data Analytics reflects the challenges of data that are too vast, too unstructured, and too fast moving to be managed by traditional methods.

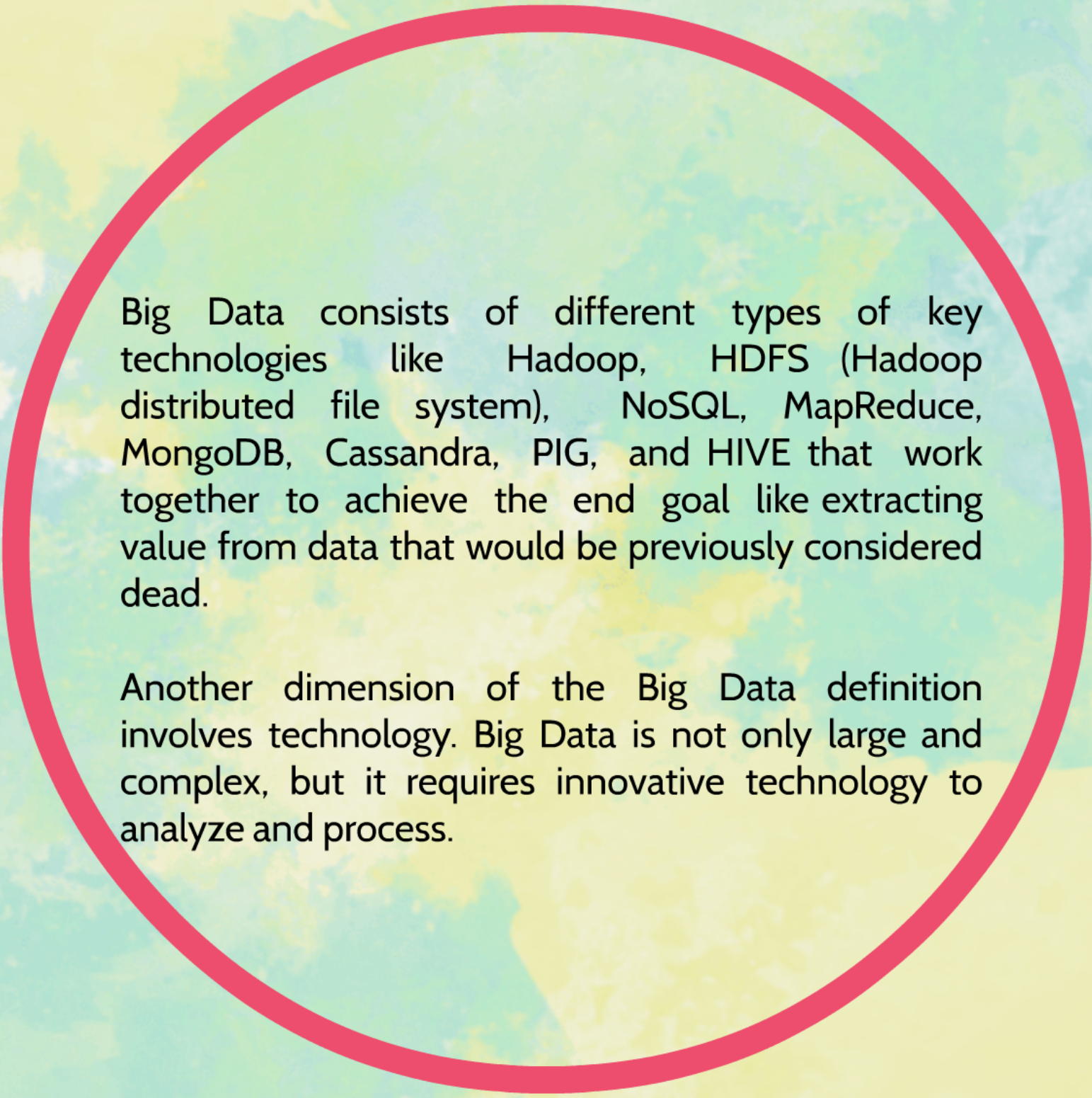
Big Data consists of different types of big data, such as structured, unstructured, and semi-structured data. It is often referred to as the 3 Vs: Volume, Velocity, and Variety.

Another dimension of the Big Data definition involves technology. Big Data is not only large and complex, but it requires innovative technology to analyze and process.



Big Data is an important concept, which is applied to data, which does not conform to the normal structure of the traditional database.

Big Data Analytics reflects the challenges of data that are too vast, too unstructured, and too fast moving to be managed by traditional methods.



Big Data consists of different types of key technologies like Hadoop, HDFS (Hadoop distributed file system), NoSQL, MapReduce, MongoDB, Cassandra, PIG, and HIVE that work together to achieve the end goal like extracting value from data that would be previously considered dead.

Another dimension of the Big Data definition involves technology. Big Data is not only large and complex, but it requires innovative technology to analyze and process.

BIG DATA TECHNOLOGIES

Apache Flume

Apache Flume is a distributed, reliable, and available system for efficiently collecting, aggregating and moving large amounts of log data from many different sources to a centralized data store.

Flume deploys as one or more agents. Agents consist of three pluggable components: sources, sinks, and channels.



Apache Sqoop

Apache Sqoop is a Command Line (CLI) tool designed to transfer data between Hadoop and relational databases.

Sqoop can import data from an RDBMS such as MySQL or Oracle Database into HDFS and then export the data back after data has been transformed using MapReduce.

Sqoop also has the ability to import data into HBase and Hive.

Apache Pig

Apache's Pig is a major project, which is lying on top of Hadoop, and provides higher-level language to use Hadoop's MapReduce library.

Pig was initially developed at Yahoo Research around 2006 but moved into the Apache Software Foundation in 2007.

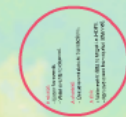
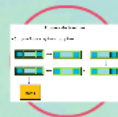
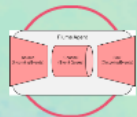
Pig provides the scripting language to describe operations like the reading, filtering and transforming, joining, and writing data which are exactly the same operations that MapReduce was originally designed for.

Instead of expressing these operations in thousands of lines of Java code which uses MapReduce directly, Apache Pig lets the users express them in a language that is not unlike a batch or Perl script.

Apache Flume

Apache Flume is a distributed, reliable, and available system for efficiently collecting, aggregating and moving large amounts of log data from many different sources to a centralized data store.

Flume deploys as one or more agents, Agents consist of three pluggable components: sources, sinks, and channels.



A source:

- Listen for events.
- Write events to channel.

A channel:

- Queue event data as transactions.

A sink:

- Write event data to target i.e. HDFS.
- Remove event from queue (channel).