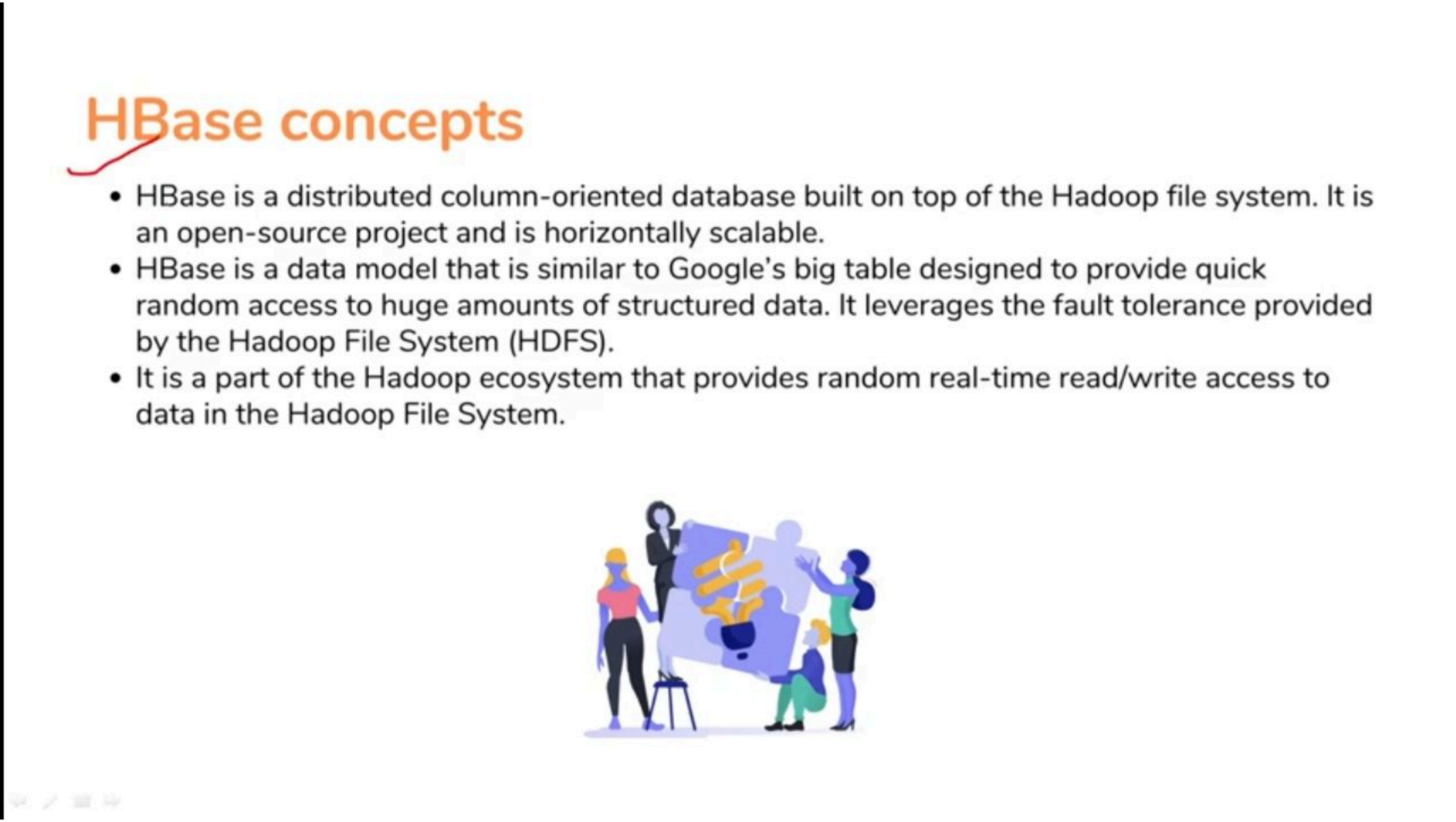
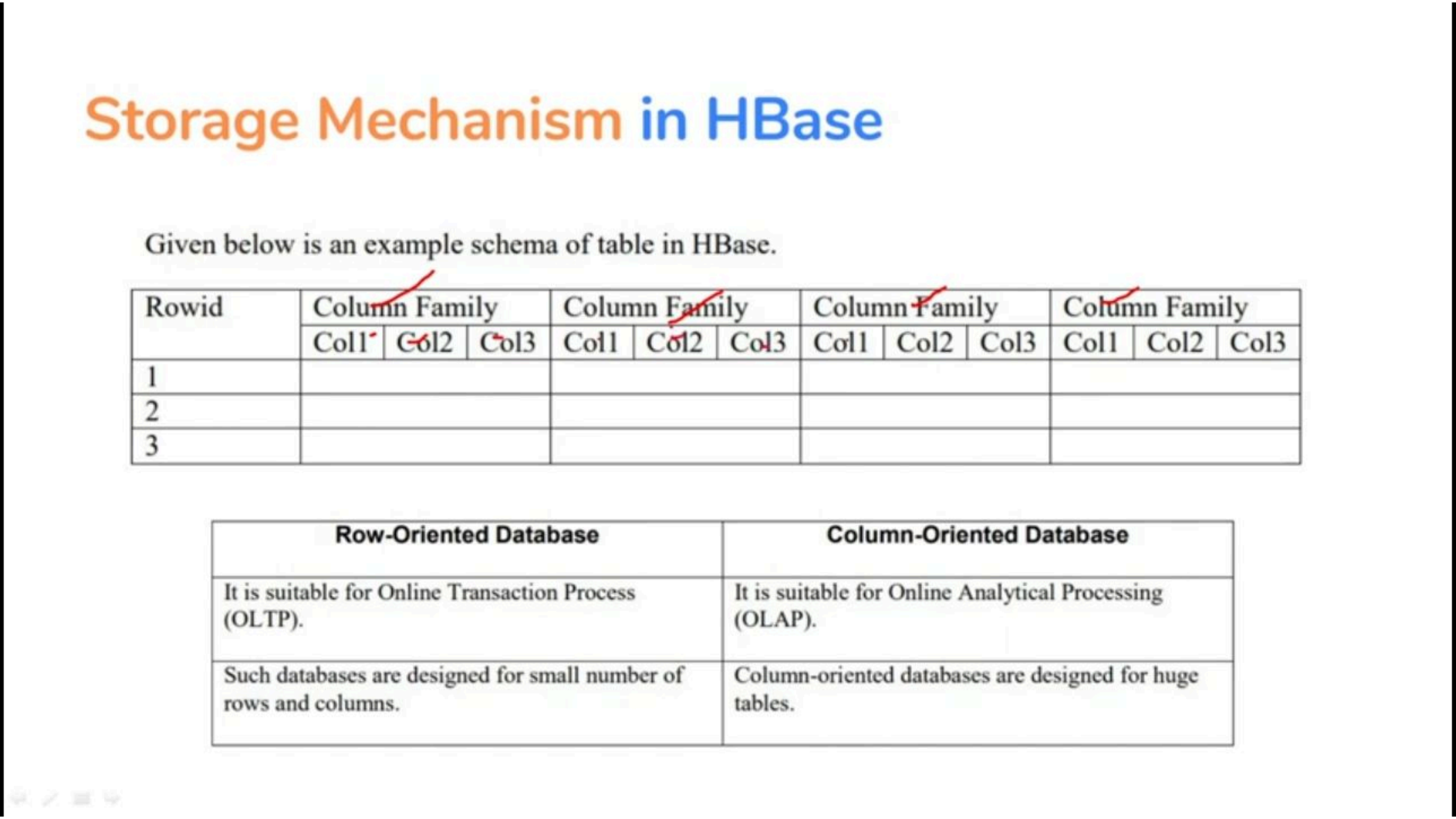


HBase concepts

- HBase is a distributed column-oriented database built on top of the Hadoop file system. It is an open-source project and is horizontally scalable.
- HBase is a data model that is similar to Google's big table designed to provide quick random access to huge amounts of structured data. It leverages the fault tolerance provided by the Hadoop File System (HDFS).
- It is a part of the Hadoop ecosystem that provides random real-time read/write access to data in the Hadoop File System.

Storage Mechanism in HBase

- HBase is a column-oriented database and the tables in it are sorted by row. The table schema defines only column families, which are the key value pairs. A table have multiple column families and each column family can have any number of columns.
- Table is a collection of rows.
- Row is a collection of column families.
- Column family is a collection of columns.
- Column is a collection of key value pairs.

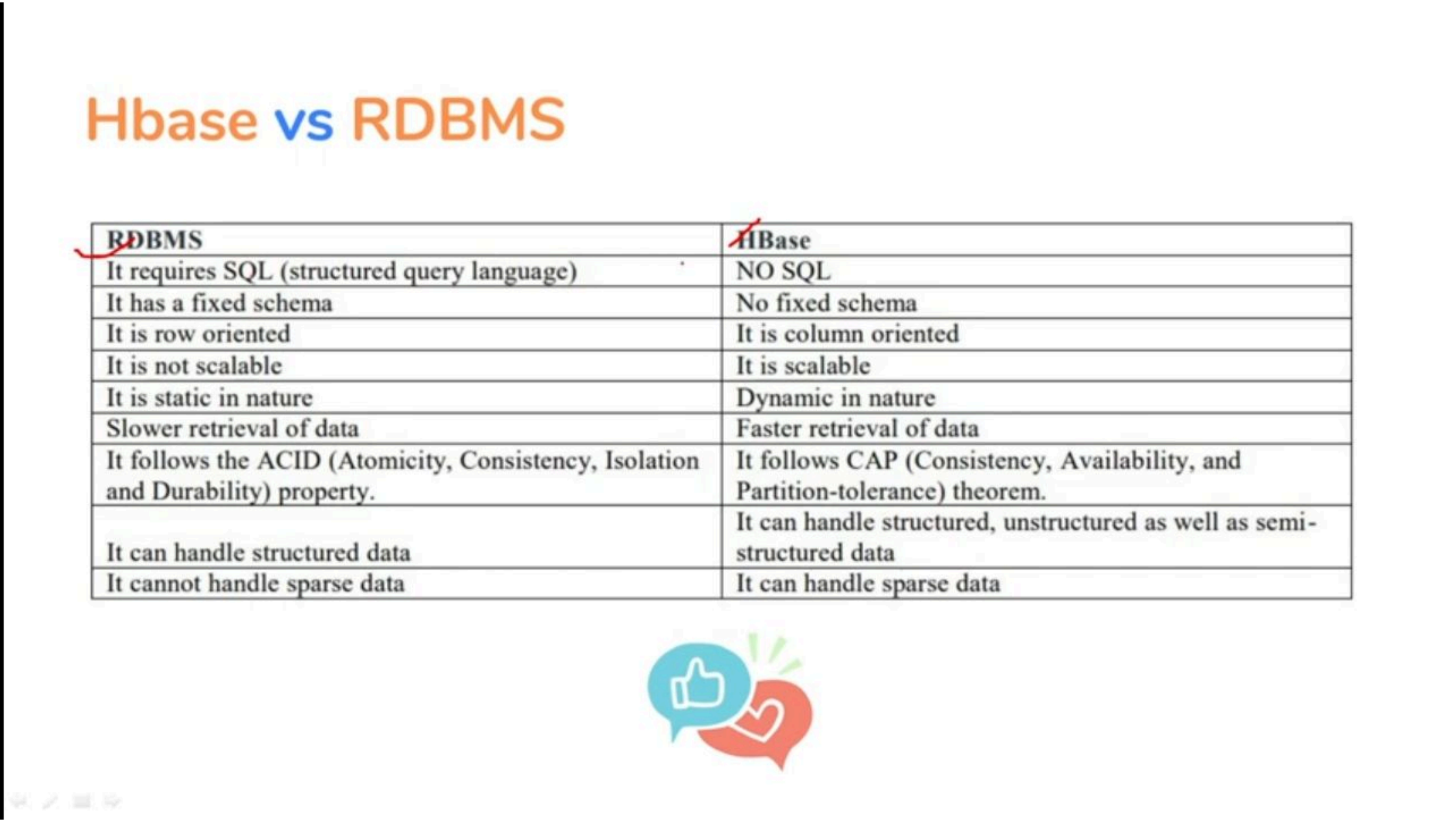


Storage Mechanism in HBase

Given below is an example schema of table in HBase.

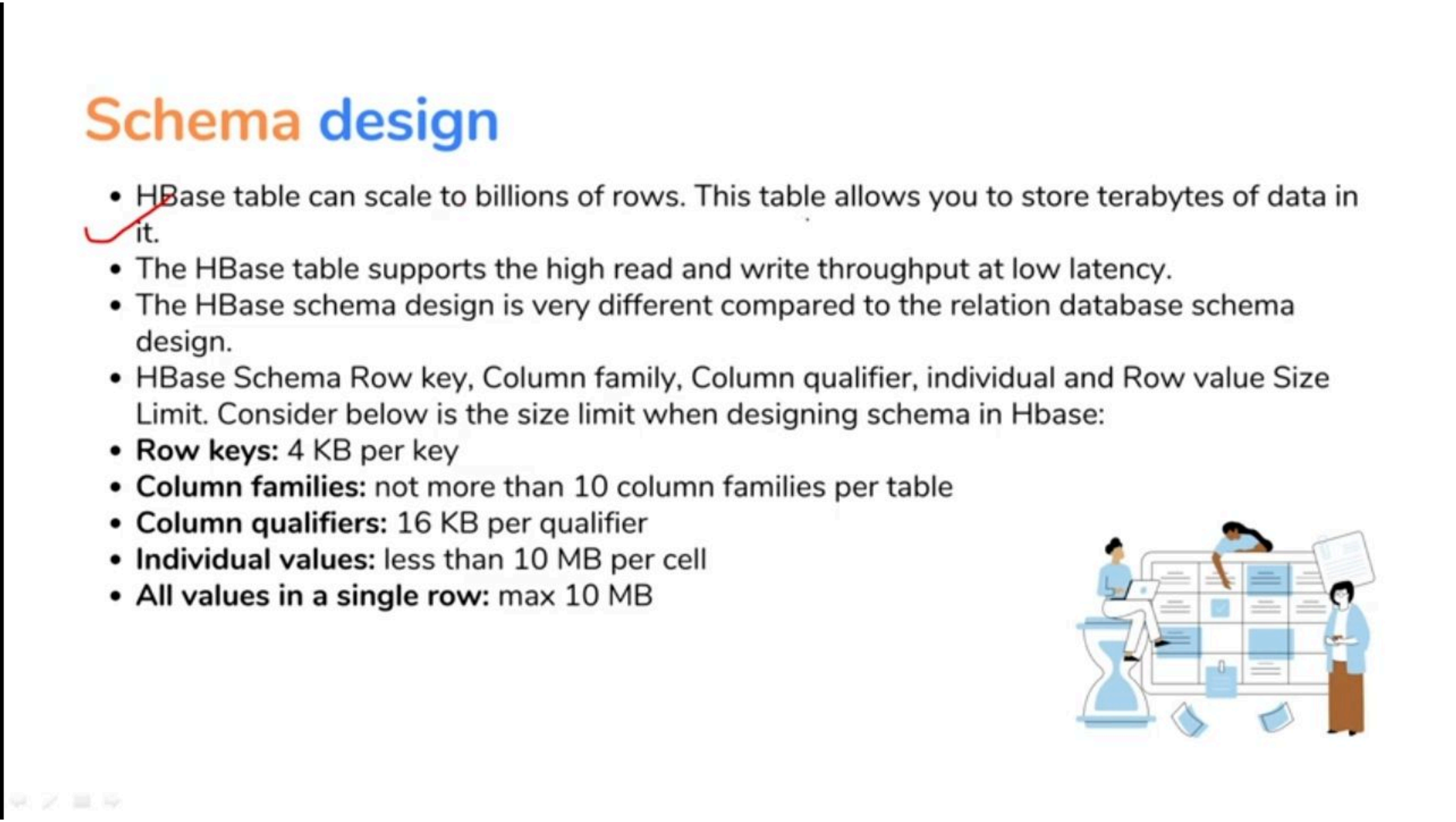
Rowid	Column Family			Column Family			Column Family			Column Family		
	Col1	Col2	Col3	Col1	Col2	Col3	Col1	Col2	Col3	Col1	Col2	Col3
1												
2												
3												

Row-Oriented Database	Column-Oriented Database
It is suitable for Online Transaction Process (OLTP).	It is suitable for Online Analytical Processing (OLAP).
Such databases are designed for small number of rows and columns.	Column-oriented databases are designed for huge tables.



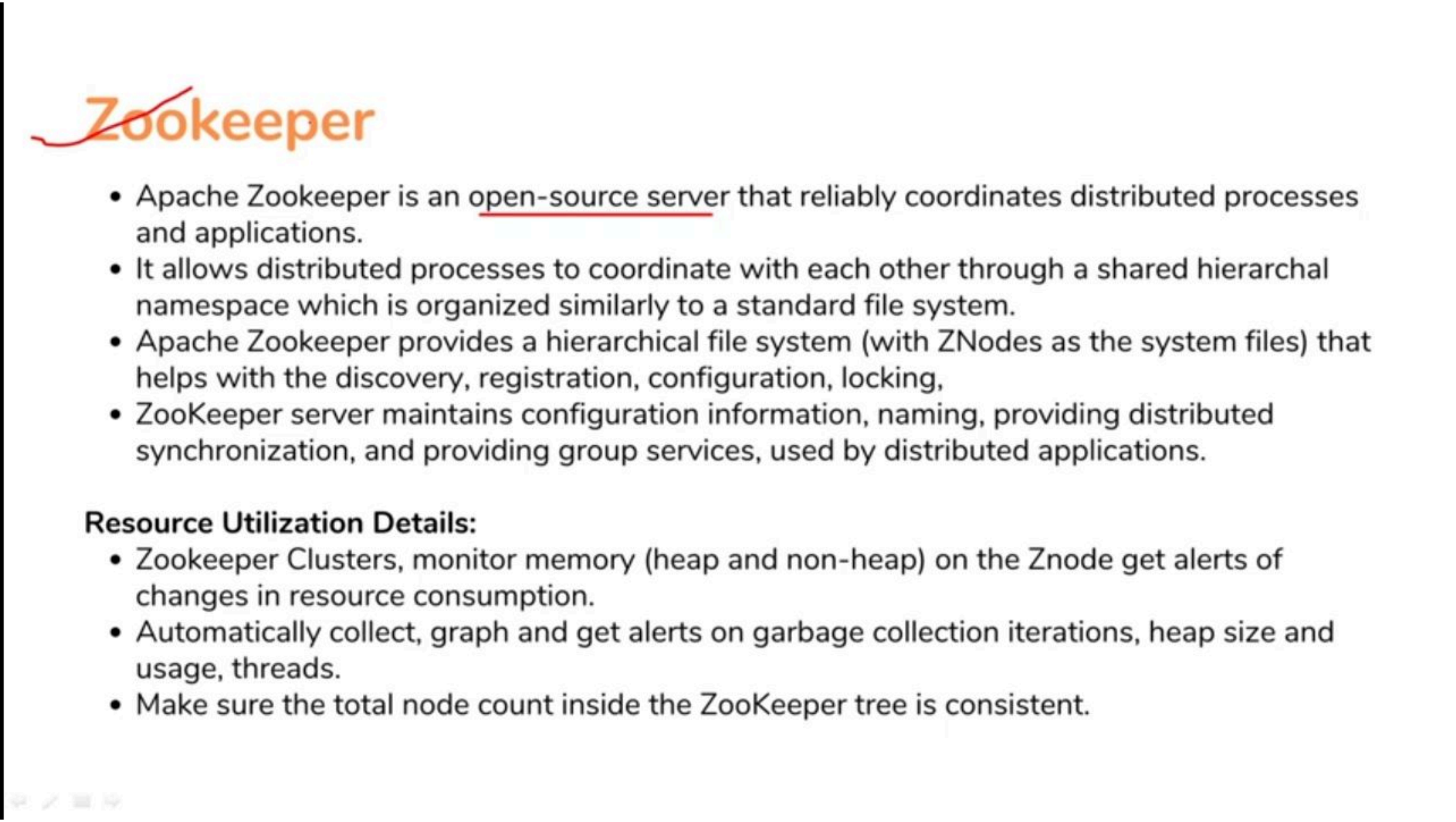
Hbase vs RDBMS

RDBMS	HBase
It requires SQL (structured query language)	NO SQL
It has a fixed schema	No fixed schema
It is row oriented	It is column oriented
It is not scalable	It is scalable
It is static in nature	Dynamic in nature
Slower retrieval of data	Faster retrieval of data
It follows the ACID (Atomicity, Consistency, Isolation and Durability) property.	It follows CAP (Consistency, Availability, and Partition-tolerance) theorem.
It can handle structured data	It can handle structured, unstructured as well as semi-structured data
It cannot handle sparse data	It can handle sparse data



Schema design

- HBase table can scale to billions of rows. This table allows you to store terabytes of data in it.
- The HBase table supports the high read and write throughput at low latency.
- The HBase schema design is very different compared to the relation database schema design.
- HBase Schema Row key, Column family, Column qualifier, individual and Row value Size Limit. Consider below is the size limit when designing schema in Hbase:
- **Row keys:** 4 KB per key
- **Column families:** not more than 10 column families per table
- **Column qualifiers:** 16 KB per qualifier
- **Individual values:** less than 10 MB per cell
- **All values in a single row:** max 10 MB

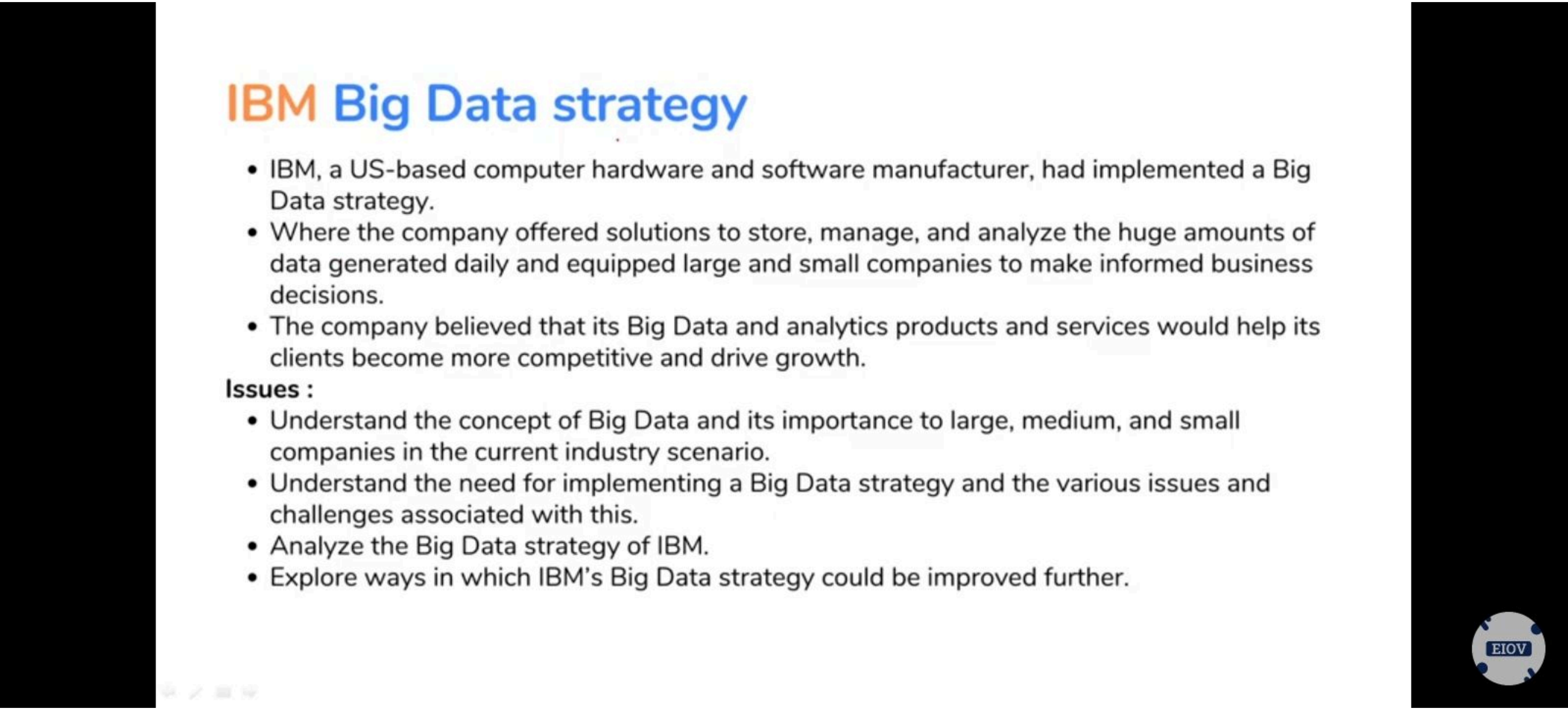


Zookeeper

- Apache Zookeeper is an open-source server that reliably coordinates distributed processes and applications.
- It allows distributed processes to coordinate with each other through a shared hierarchal namespace which is organized similarly to a standard file system.
- Apache Zookeeper provides a hierarchical file system (with ZNodes as the system files) that helps with the discovery, registration, configuration, locking,
- ZooKeeper server maintains configuration information, naming, providing distributed synchronization, and providing group services, used by distributed applications.

Resource Utilization Details:

- Zookeeper Clusters, monitor memory (heap and non-heap) on the Znode get alerts of changes in resource consumption.
- Automatically collect, graph and get alerts on garbage collection iterations, heap size and usage, threads.
- Make sure the total node count inside the ZooKeeper tree is consistent.

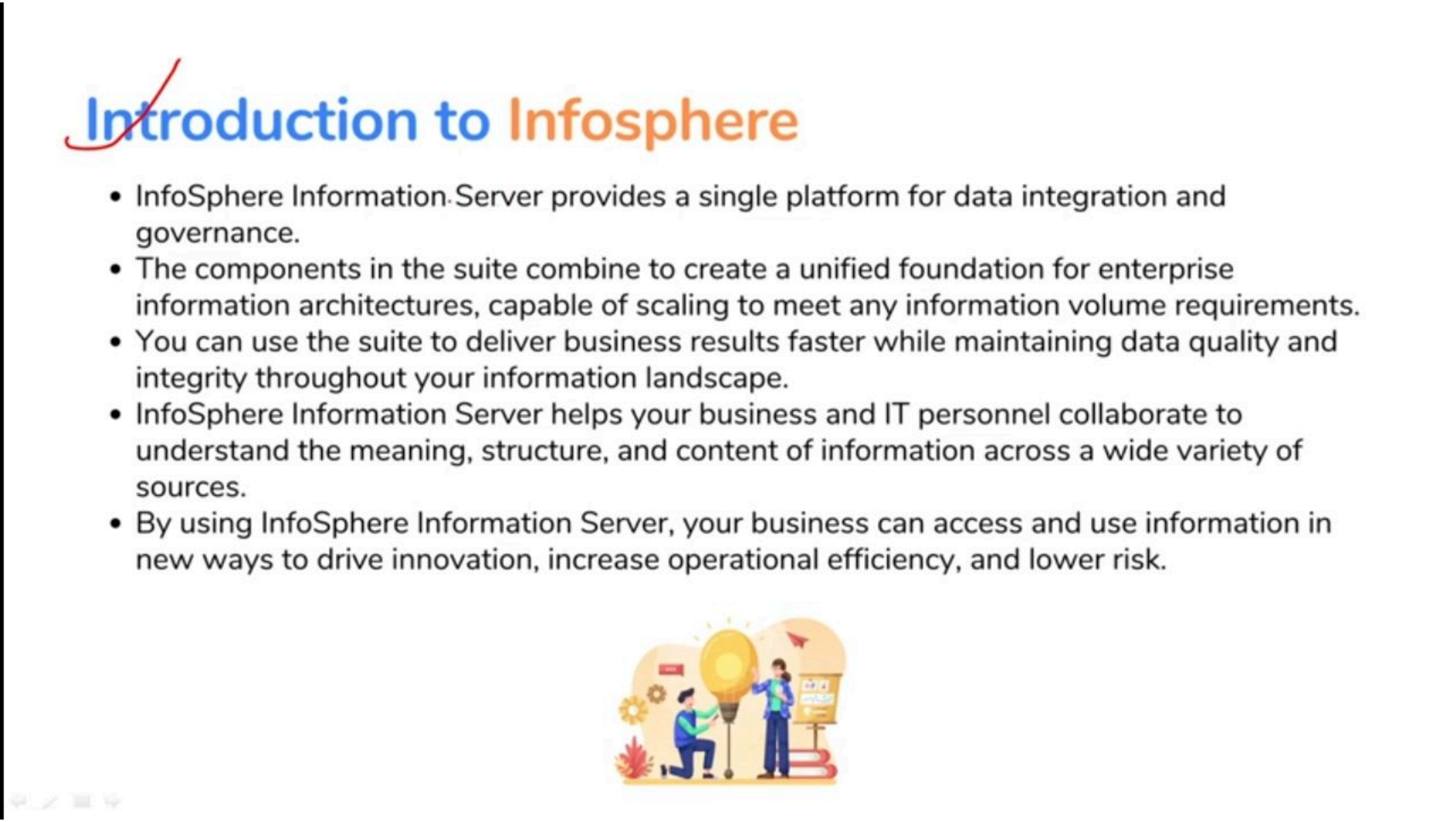


IBM Big Data strategy

- IBM, a US-based computer hardware and software manufacturer, had implemented a Big Data strategy.
- Where the company offered solutions to store, manage, and analyze the huge amounts of data generated daily and equipped large and small companies to make informed business decisions.
- The company believed that its Big Data and analytics products and services would help its clients become more competitive and drive growth.

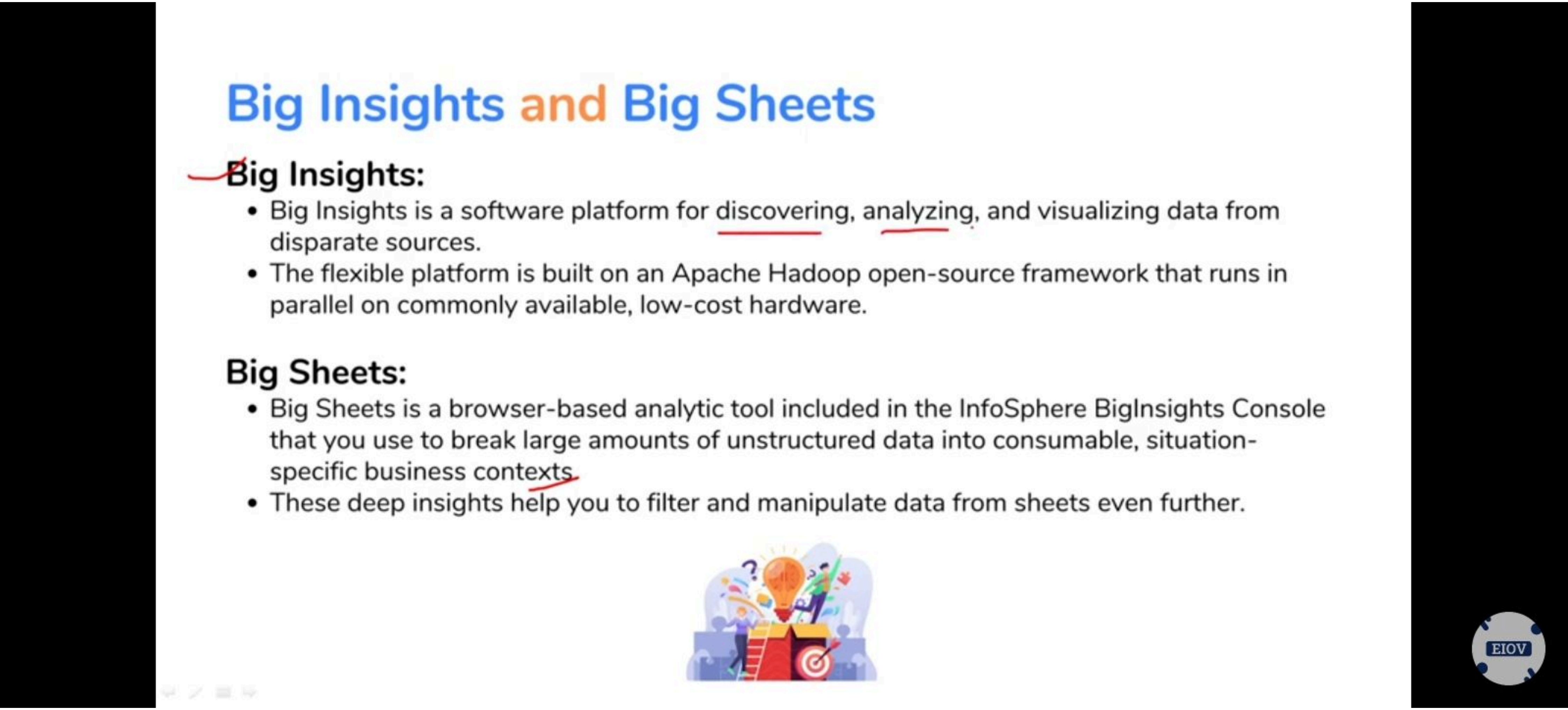
Issues :

- Understand the concept of Big Data and its importance to large, medium, and small companies in the current industry scenario.
- Understand the need for implementing a Big Data strategy and the various issues and challenges associated with this.
- Analyze the Big Data strategy of IBM.
- Explore ways in which IBM's Big Data strategy could be improved further.



Introduction to Infosphere

- InfoSphere Information Server provides a single platform for data integration and governance.
- The components in the suite combine to create a unified foundation for enterprise information architectures, capable of scaling to meet any information volume requirements.
- You can use the suite to deliver business results faster while maintaining data quality and integrity throughout your information landscape.
- InfoSphere Information Server helps your business and IT personnel collaborate to understand the meaning, structure, and content of information across a wide variety of sources.
- By using InfoSphere Information Server, your business can access and use information in new ways to drive innovation, increase operational efficiency, and lower risk.



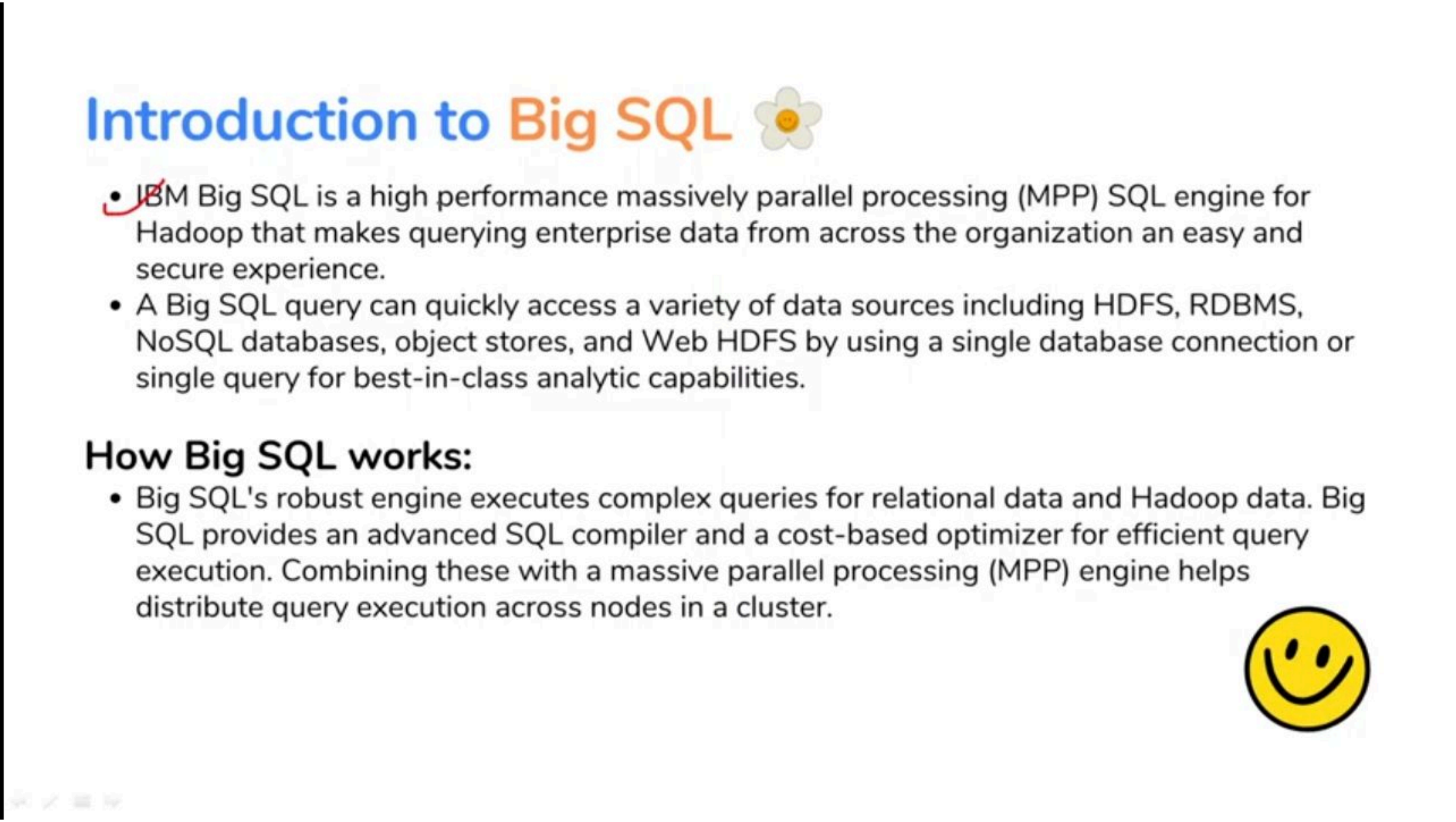
Big Insights and Big Sheets

Big Insights:

- Big Insights is a software platform for discovering, analyzing, and visualizing data from disparate sources.
- The flexible platform is built on an Apache Hadoop open-source framework that runs in parallel on commonly available, low-cost hardware.

Big Sheets:

- Big Sheets is a browser-based analytic tool included in the InfoSphere BigInsights Console that you use to break large amounts of unstructured data into consumable, situation-specific business contexts.
- These deep insights help you to filter and manipulate data from sheets even further.



Introduction to Big SQL

- IBM Big SQL is a high performance massively parallel processing (MPP) SQL engine for Hadoop that makes querying enterprise data from across the organization an easy and secure experience.
- A Big SQL query can quickly access a variety of data sources including HDFS, RDBMS, NoSQL databases, object stores, and Web HDFS by using a single database connection or single query for best-in-class analytic capabilities.

How Big SQL works:

- Big SQL's robust engine executes complex queries for relational data and Hadoop data. Big SQL provides an advanced SQL compiler and a cost-based optimizer for efficient query execution. Combining these with a massive parallel processing (MPP) engine helps distribute query execution across nodes in a cluster.