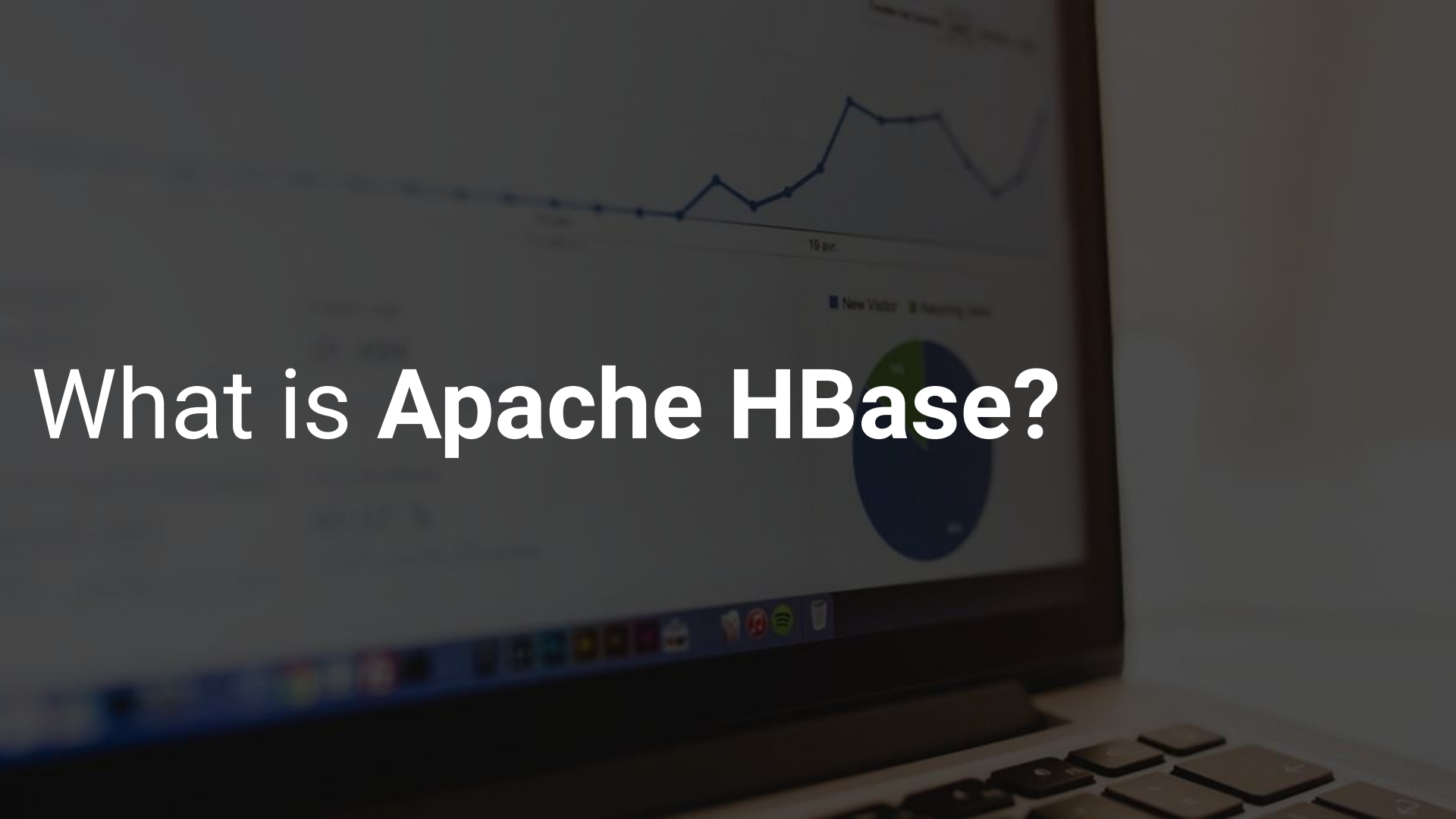
APACHE

HBASE



What is **Apache HBase?**

HBase is a Hadoop project which is Open Source, distributed Hadoop database which has its genesis in the Google's Bigtable.

- Its programming language is Java.
- Now, it is an integral part of the Apache Software Foundation and the Hadoop ecosystem.
- Also, it is a high availability database which exclusively runs on top of the HDFS.
- It is a column-oriented database built on top of HDFS.

Why Apache HBase?

HBase features like working with sparse data in an extremely fault-tolerant and resilient way and the way it can work on multiple types of data also making it useful for varied business scenarios.

Why should you use HBase?

Along with HDFS and MapReduce, **HBase** is one of the core components of the Hadoop ecosystem. Here are some salient features of HBase which make it significant to use:

- Apache HBase has a completely distributed architecture.
- It can easily work on extremely large scale data.
- HBase offers high security and easy management which results in unprecedented high write throughput.
- For both structured and semi-structured data types we can use it.
- Moreover, the MapReduce jobs can be backed with HBase Tables.

Apache HBase Architecture?

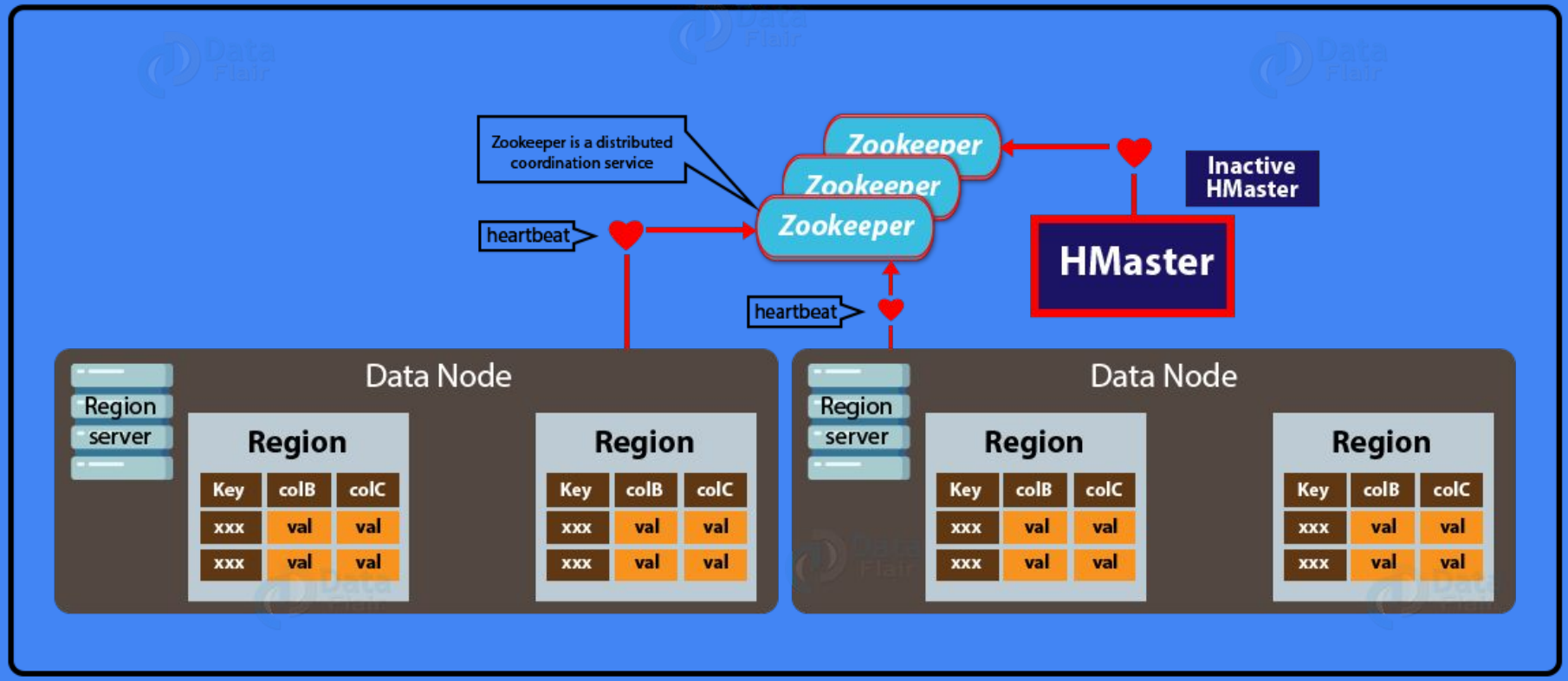
HBase Architecture is basically a column-oriented key-value data store and also it is the natural fit for deploying as a top layer on HDFS because it works extremely fine with the kind of data that Hadoop process.

Moreover, when it comes to both read and write operations it is extremely fast and even it does not lose this extremely important quality with humongous datasets.

There are 3 major components of HBase Architecture:

- Zookeeper
- HMaster server
- Region servers

Zookeeper


Zookeeper is a distributed coordination service
Zookeeper
Zookeeper
Zookeeper
Inactive HMaster
HMaster
heartbeat
heartbeat
Data Node
Region server
Region
Key
colB
colC
xxx
val
val
Region
Key
colB
colC
xxx
val
val
Data Node
Region server
Region
Key
colB
colC
xxx
val
val
Region
Key
colB
colC
xxx
val
val

HBase Storage mechanism

Basically, **HBase** is a column-oriented database.
Moreover, the tables in it are sorted by row.
Here, the table schema defines only column families, which are the key-value pairs. However, it is possible that a table has multiple column families and here each column family can have any number of columns.
Moreover, here on the disk, subsequent column values are stored contiguously.
And, also each cell value of the table has a timestamp here.

HBase Performance Tuning

- The table refers to the collection of rows.
- Row refers to the collection of column families.
- Column family refers to the collection of columns.
- The column refers to the collection of key-value pairs.

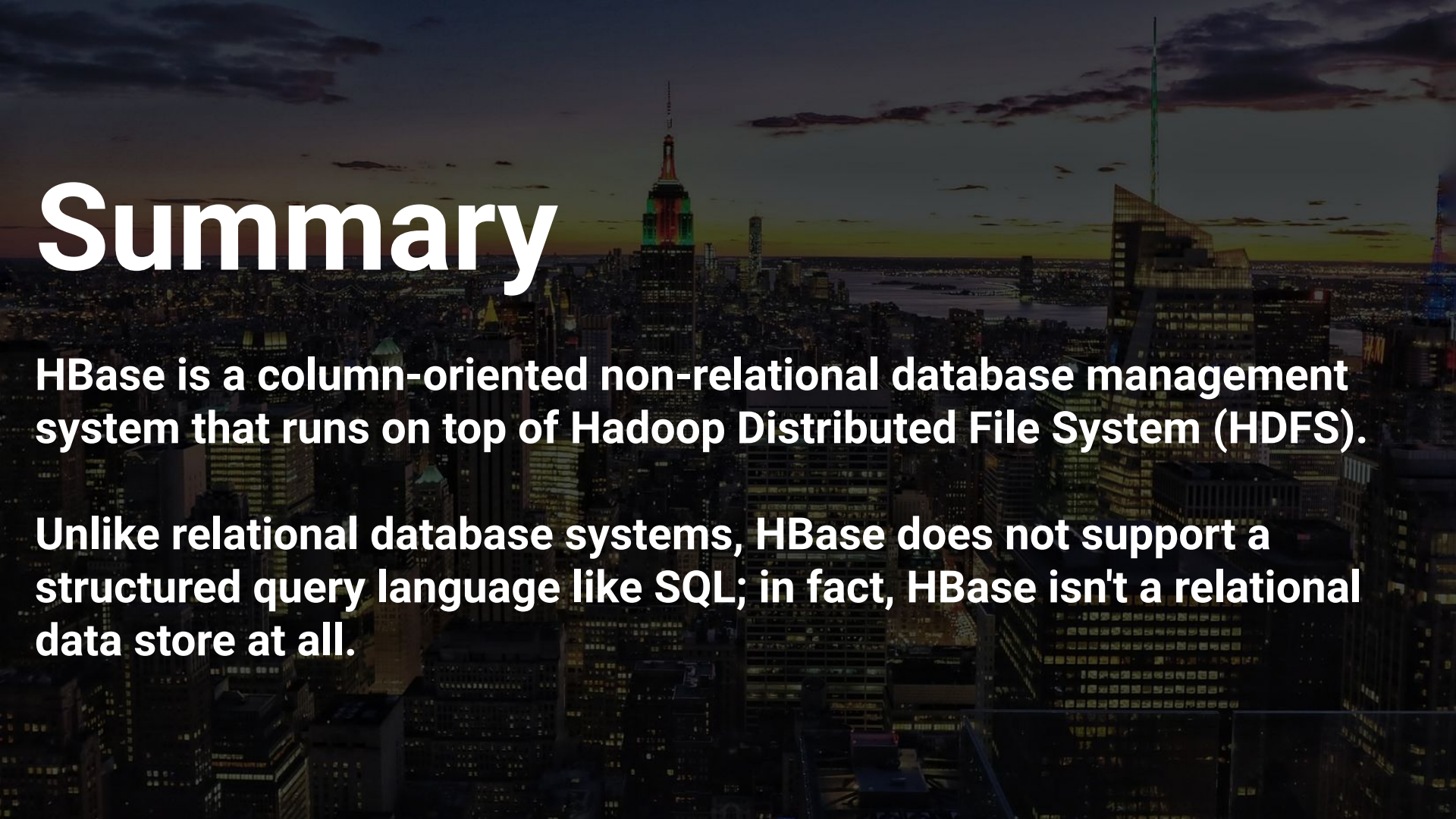
Databases in **HBase** which store data tables as sections of columns of data, instead of rows of data are Column-oriented Databases.
In simple words, they will have column families.

HBase Features

- Apache HBase is linearly scalable.
- Moreover, it provides automatic failure support.
- It also offers consistent read and writes.
- We can integrate it with Hadoop, both as a source as well as the destination.
- Also, it has easy java API for the client.
- HBase also offers data replication across clusters.

HBase Use Cases

- While we want to have random, real-time read/write access to Big Data, we use Apache HBase.
- It is possible to host very large tables on top of clusters of commodity hardware with Apache HBase.
- After Google's Bigtable, HBase is a non-relational database modeled. Basically, as Bigtable acts up on Google File System, in same way HBase works on top of Hadoop and HDFS.



Summary

HBase is a column-oriented non-relational database management system that runs on top of Hadoop Distributed File System (HDFS).

Unlike relational database systems, HBase does not support a structured query language like SQL; in fact, HBase isn't a relational data store at all.