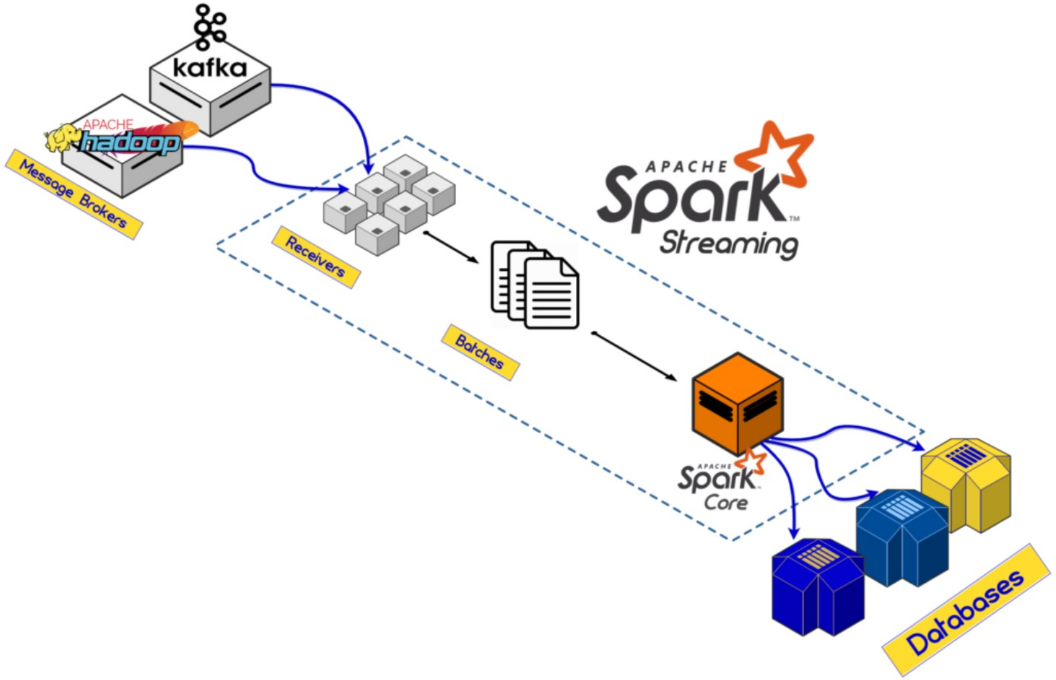
Hadoop & Spark

- Hadoop is the foundation of your big data architecture
  - HDFS is a distributed file system that handles large data sets running on commodity hardware. It is used to scale a single Apache Hadoop cluster to hundreds (and even thousands) of nodes.
- It is responsible for storing and processing your data
- Spark is an in-memory processing engine that can perform real-time stream processing or batch processing on data stored in Hadoop

Apache Kafka is a distributed streaming platform that allows developers to create applications that continuously produce and consume data streams. As such, it enables the creation of applications that react to events as they happen in real-time.



Apache Spark is a general-purpose distributed processing system used for big data workloads that provides high-level APIs in Java, Svala, Python and R. It was designed to replace MapReduce and improve upon its shortcomings, such as slow batch processing times and lack of support for interactive real-time data analysis.