Damon Quire

Hadoop is not a database. It is a programming framework that is used to process and store extremely large data sets across multiple computers. It has two basic components: the Hadoop Distributed File System which is used for storage and the MapReduce engine which is used to transform it and query the data. The main difference between it and some of the other databases like SQL and Oracle that we’ve went over is basically that first sentence. It’s simply not a database. It’s a distributed file system. Which means it lets you store and process large amounts of data on a cloud that handles data redundancy. It doesn’t process queries immediately like SQL because of the massive amount of data. It uses what’s called Map-Reduce in order to do this. It processes info on each node instead of processing it all by moving it to one network which can lead to overloads of data. In SQL when moving data from one database to another it’s required that you know it’s metadata, it’s structure, type, etc. It will error if it doesn’t follow these guidelines established by the desired database destination. This is called Schema on write. Hadoop uses Schema on Read which means that you can actually just bring in the info to the desired database without needing to know or following the structure of the destination. However, when you want to read the data and actually make it make sense to the human eye, you then have to put in code the structure you want it to appear in. This makes it to where you don’t have to preprogram in the structure of the database before inserting the info into it. Another key difference is how the data is stored. Relational databases store their info in tables defined by schema, whereas Hadoop uses what are called key value pairs. The main difference however is what I began with: Hadoop is simply NOT a database.