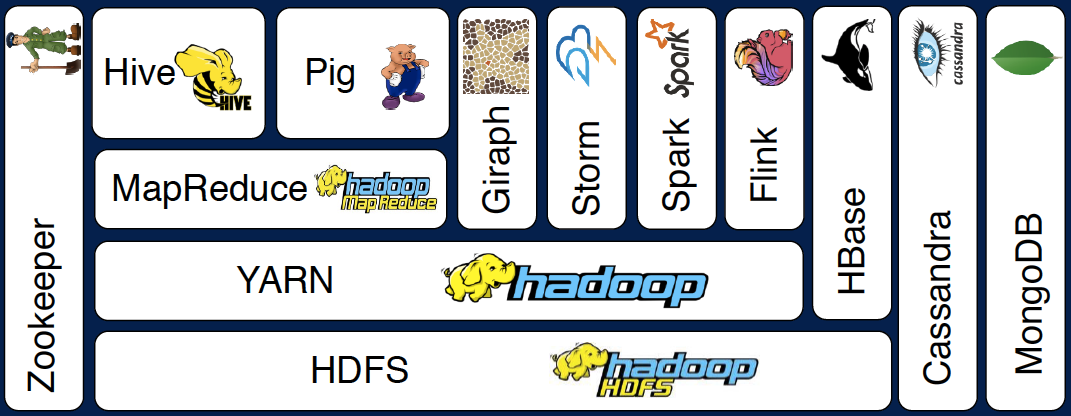
**Overview**

* Used to be primarily for batch processing but not anymore.
  + Interactive environments
  + Web apps built with a Hadoop server backend that allows fast transactions
* Horizontal scaling
  + Linear scaling
  + Can scale out with less, if any, diminishing marginal rate of return
  + More efficient than vertical scaling in a lot of ways

-------------------------------------------------------------------------------------------------------------------------------

****

**Core Hadoop Ecosystem**

* HDFS
  + Hadoop Distributed File System
    - File system that allows the distribution of data across multiple file systems
    - Maintains redundant copies of data
      * Recover from crashes and unexpected outages
* YARN
  + Yet Another Resource Negotiator
  + Manages the resources on the computing cluster
    - Task management
    - Node availability monitoring
    - Heartbeat of a cluster
  + Allows us to build on top and take advantage of this optimized cluster manager (i.e. MapReduce)
* MapReduce
  + Programming model that allows us to process data across a cluster
  + Mappers & Reducers
    - Mappers allows us to efficiently transform our data in parallel across the computing cluster
    - Reducers aggregates the data mapped through Mappers
  + MapReduce and YARN used to be the same thing but as they were separated recently, this allowed for extensions to be built on top of YARN that solves
* Pig
  + High level programming API that allows users to script SQL-like queries to gain complex answers
  + Transforms these scripts into a process that can run on MapReduce
  + High level scripting language sitting on top of MapReduce
* Hive
  + An interface that allows users to interact with the distributed database (cluster) as if it were one database
  + The engine that allows SQL queries to be run on the Hadoop environment
  + Transforms SQL queries to distributed queries that query the cluster
* Ambari
  + A Dashboard for the visualization of a cluster
  + Allows users to monitor the cluster and use GUIs to execute functionality on the cluster
  + Alternatives exist
* MESOS
  + Alternative to YARN
  + Resource management on the cluster
* Spark
  + On top of Yarn or Mesos
  + Allows efficient and fast data processing on clusters
  + Write Spark scripts with Scala, Python, or Java
  + Reliable and versatile
    - Handles streaming data in real-time
    - Enables machine learning over a cluster
    - Handles SQL queries
  + DAGs
* TEZ
  + Uses similar techniques as Spark
  + DAGs
    - Allows to produce more optimal query plans
  + Alternative to MapReduce
* Apache HBase
  + A “NoSQL database”
  + Fast way of exposing data stored in a Hadoop cluster to applications
* Apache Storm
  + A way of processing streaming data quickly and in real-time
  + Similar to Spark in allowing data streaming but in a different way
  + Not a batch thing anymore
    - Can update machine learning models or data in real-time
* Oozie
  + A job scheduler for a cluster
  + Schedules a list of tasks involving multiple steps and systems
  + Similar to SSIS for SQL Server?
  + Adds reliability and consistency to managing complex task chains
* Zookeeper
  + Monitor node status
  + Mainly used for coordinating everything on a cluster
  + Many applications rely on Zookeeper to maintain reliable and consistent performance in case a node collapses
* Scoop
  + Connector between HDFS and legacy databases
  + Ties the Hadoop cluster with relational databases
* Flume
* Kafka

**External Data Storage**

* MySQL
  + Spark and similar technology can use a relational database such as MySQL to store and retrieve data from
* Cassandra
* mongoDB
* HBase (part of the Hadoop stack)
* NoSQL databases such as Cassandra or mongoDB is a good layer between a real-time application and a cluster to efficiently manage data

**Query Engines**

* Apache Drill
* Hue
  + Takes the role of Ambari for Cloudera
* Apache Phoenix
  + Similar to Drill
  + Takes it a step further than drill by giving you ACID guarentees and OLTP
  + Can make a the cluster look a lot like a relational datastore
* Presto
* Apache Zeppelin
  + A “notebook” approach in the UI that allows interactions with the cluster
* Apache Hive (part of the Hadoop stack)