Big Data

- Data ingestion (kafka)
  - Collects data from a cluster/web server and broadcast it into Hadoop cluster
- Spark
  - Same level of map reduce and sits on top of yarn
  - Spark scripts are written in Scala java python
  - Efficiently processes data on Hadoop cluster
  - Handles sql queries machine learning streaming data
- Hadoop
  - Software platform for distributed storage and distributed processing of very large data sets scaled horizontally
  - HDFS (Hadoop distributed file system)
    - System that allows you to distribute data across clusters to make it look like one giant file system
    - keeps redundant copy
  - YARN
    - Yet another resource negotiator
    - Data processor
    - manages resources on cluster
    - Heatbeat
  - MAPREDUCE
    - Programming model that allows you to process data across a cluster
    - Consists of mappers and reducers
    - Mappers transforms data in parallel across cluster efficiently
    - Reducers aggregate data together
  - HIVE
    - Takes sql queries and makes distributed files look like a sql database
    - On top of map reduce
- Storm
  - Way of processing streaming data in real time
  - Spark streaming is similar storm does it differently
- Mysql, cassandra, mongodb
  - Exposes data for realtime usage
  - Sits between realtime apps and clusters
- Apache hbase
  - Hbase is a nosql database
  - exposes that on cluster
- Redis
  - In memory key/value data store
  - Good for frequently updated real time data

What is a data pipeline?

- Efficient flow of data from one location to the other
- Useful analysis cannot begin until data is available

What is an ETL pipeline

- Extract, transform, load
- Extracts data from one system transforms the data and loss it into a database
- Subset of data pipeline