1. Thing Big
   1. High Performance Computing
2. Yankee Group
   1. 1970’s
3. Day 1
   1. Intro to Big Data
   2. Cluster Architecture
   3. SQL on Hadoop
   4. Big Data Bottlenecks
   5. Spark Architecture
   6. Spark in Scala
4. A quick introduction to big data
   1. Tesla
   2. 3 V’s
      1. Volume
      2. Variety
      3. Velocity
   3. Big data – anything that doesn’t fit on your laptop
5. Hadoop
   1. Commodity Hardware
   2. Spend resources to save resources
   3. Send code where the data lives
      1. JVM – you send your 1gb < programs to where the data lives
   4. Interpret data when used (schema on read)
   5. Compute results only on demand
6. Data can be in 4 places
   1. Local
      1. Users/chowe/localfile
      2. scp
   2. Edge Node
      1. /home/Hadoop/Masterfile
      2. hdfs
   3. HDFS
      1. Move to hive tables
         1. Create table
      2. /user/Hadoop/hdfsfile/
   4. Hive Tables
      1. /user /hive/warehouse/table
7. Terminal commands
   1. Yarn
      1. Run java program jarfile with arguments args
      2. Yarn jar jarfile args
8. Commands in Yarn
   1. Cluster – help
      1. Helps you find out information about the nodes
9. Command line same interface
   1. Master node
      1. <http://ec2-54-221-189-104.compute-1.amazonaws.com:8088/cluster>
   2. Hadoop nodes
      1. <http://ec2-54-221-189-104.compute-1.amazonaws.com:50070/dfshealth.html#tab-overview>
   3. Local directory vs hdfs vs dfs
   4. <http://ec2-54-221-189-104.compute-1.amazonaws.com:50070/explorer.html#/>
10. Schema on read
    1. Interpret files in any way that we like
    2. Hive
       1. Sql based tool for data warehousing using Hadoop clusters
       2. Lower barrier for Hadoop adoption for existing sql apps
       3. Hive is not a database
       4. Schema on read engine
       5. Programming Hive
          1. Dean wampler
       6. You can define table schemas
    3. Table storage
    4. Elastic Map Reduce
    5. Metastore
       1. Retain metadata
          1. Store metadata
11. External Tables
    1. A way of allowing hive to do this for us
    2. Partitioning
       1. Hadoop
    3. Creating partitions by dates
    4. Run the partitions first and then the where
12. Does not support Amazon S3
13. Inserting Data
    1. Dynamic partition inserts
       1. Partition (<Field 1>, <Field 2>, <Field 3>)
14. Properties
    1. Hive.exec.dynamic.partition=true
    2. Set hive.exec.dynamic.partition.mode=dynamic
15. File and Record Format
    1. Sequence files
       1. Binary format files
    2. RCLife
       1. Columner format
16. Custom File Format
    1. Write your own in java
    2. Input and output format
17. Summary
    1. Hive advantages
       1. Provides the hive warehouse for creating cluster wide tables
       2. Indepensable for users with sql
       3. 3rd party analyst tools
18. Big Data Bottlenecks
    1. Clusters must minimize disk use to be fast
    2. Trick is allowing programmer to minimize the bottlenecks
19. Rules of good cluster
    1. Ingest into HDFS first using as parallel an operation as possible
       1. Teradata
    2. Do sanity checks. Understand how long it takes to ingest data
    3. Use in memory tools
    4. Turn big data into small data as soon as possible
    5. Focus exports on small data, not big
20. ?
    1. Can you partition a table differently than the default
21. Spark Architecture Literature
    1. Learning Spark
    2. Scala Cookbook
    3. Machine Learning with Spark
    4. Online
       1. spark.apache.org/docs/latest
       2. Primary place to get the information on spark
22. Architecture
23. Overview for the next days
    1. Spark in Scala
    2. Spark in SQL
    3. Spark in Python
       1. Very similar to Spark in Scala
    4. Spark in R
       1. Completely different
       2. Spark SQL

# September 6, 2017 – Wednesday

## 9:00 AM Recap

1. What makes Hadoop different

# September 7, 2017 – Thursday

## 9:00 AM

1. Hive – invented by FB and became open source
2. Spark ML
3. <http://ec2-54-221-189-104.compute-1.amazonaws.com:8787/auth-sign-in>

## Amazon Web Service

1. Amazon Account
2. Analytics
   1. EMR
      1. Clusters
3. Create Clusters
   1. Root device
   2. 10 GB
4. Amazon pricing on demand
   1. M3.2xlarge
5. Bootstrapping actions
6. Run some scripts to install packages and applications
7. Install Amazon Command Line stuff

## Spark Streaming

1. Ganglia
2. Cloudera – a viewer version of the course
3. Streaming
   1. Scala
   2. Java
4. Structured Streaming Pluses and Minuses