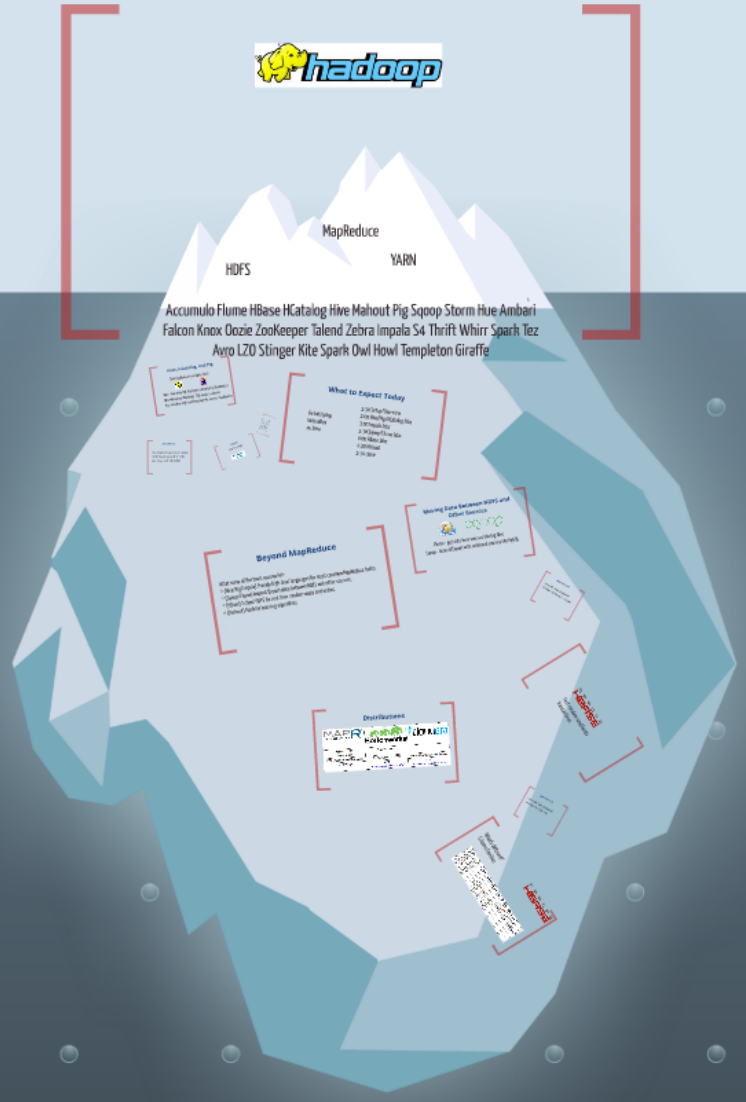
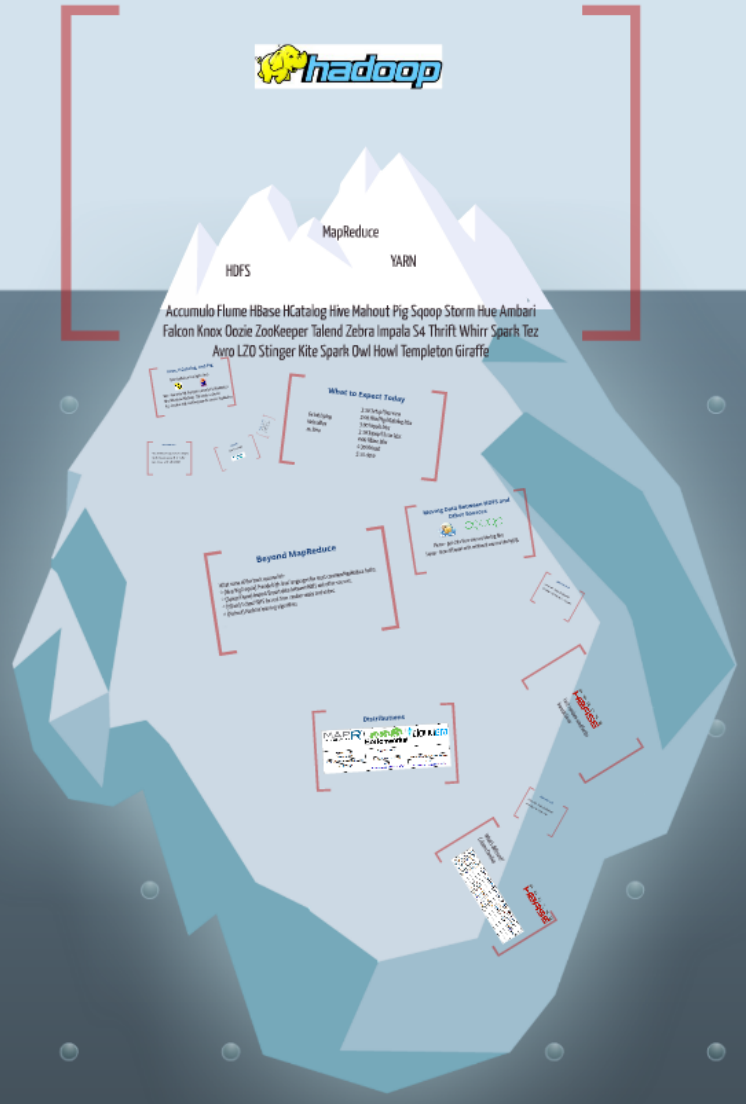


Hadoop's Friends
A Big Data Arsenal

To Start You'll Need:
A 64 bit laptop
Install VirtualBox Now!



Hadoop's Friends
A Big Data Arsenal

To Start You'll Need:
A 64 bit laptop
Install VirtualBox Now!

hadoop
MapReduce
HDFS
YARN
Accumulo Flume HBase HCatalog Hive Mahout Pig Sqoop Storm Hue Ambari
Falcon Knox Oozie ZooKeeper Talend Zebra Impala S4 Thrift Whirr Spark Tez
Avro LZO Stinger Kite Spark Owl Howl Templeton Giraffe
Hive, HCatalog, and Pig
Take MapReduce to a higher level
What to Expect

Beyond MapReduce

What some of the tools accomplish:

* (Hive/Pig/Impala) Provide high level languages for most common MapReduce tasks.
* (Sqoop/Flume) Import/Export data between HDFS and other sources.
* (HBase) Extend HDFS for real time random reads and writes.
* (Mahout) Machine learning algorithms

What to Expect Today

64 bit laptop
VirtualBox
no Java

1:30 Setup/Overview
2:00 Hive/Pig/HCatalog labs
3:00 Impala labs
3:30 Sqoop/Flume labs
4:00 HBase labs
4:30 Mahout
5:30 close

Distributions

MAPR TECHNOLOGIES	Hortonworks	cloudera
Amazon EMR Google Cloud	Windows Azure	
RHEL/CentOS/SUSE/Ubuntu	RHEL/CentOS/SUSE	RHEL/CentOS/Ubuntu/Debian
M3/M5/M7	HDP (Hortonworks Sandbox VM)	CDH (Cloudera Quickstart VM)

Hive, HCatalog, and Pig

Take MapReduce to a higher level

Hive - You write SQL that gets converted to MapReduce

Hive Metastore/HCatalog - SQL needs a schema

Pig - Another high level language for simpler MapReduce

Labs until 3:00

Stop all other local Hadoop processes (check jps)

Start Hortonworks Sandbox VM in VirtualBox

Open a browser set to localhost:8888/

Impala

SQL without MapReduce

Impala Lab until 3:30

Shut down Hortonworks Sandbox

Start up Cloudera Quickstart VM

Follow PDF lab file provided by instructor

Moving Data Between HDFS and Other Sources

Flume - get data from sources like log files

Sqoop - Import/Export with relational sources like MySQL

Labs until 3:30

Continue with Cloudera Quickstart VM

Follow Sqoop and Flume Labs in PDF provided

Fast random read/write

Transactions

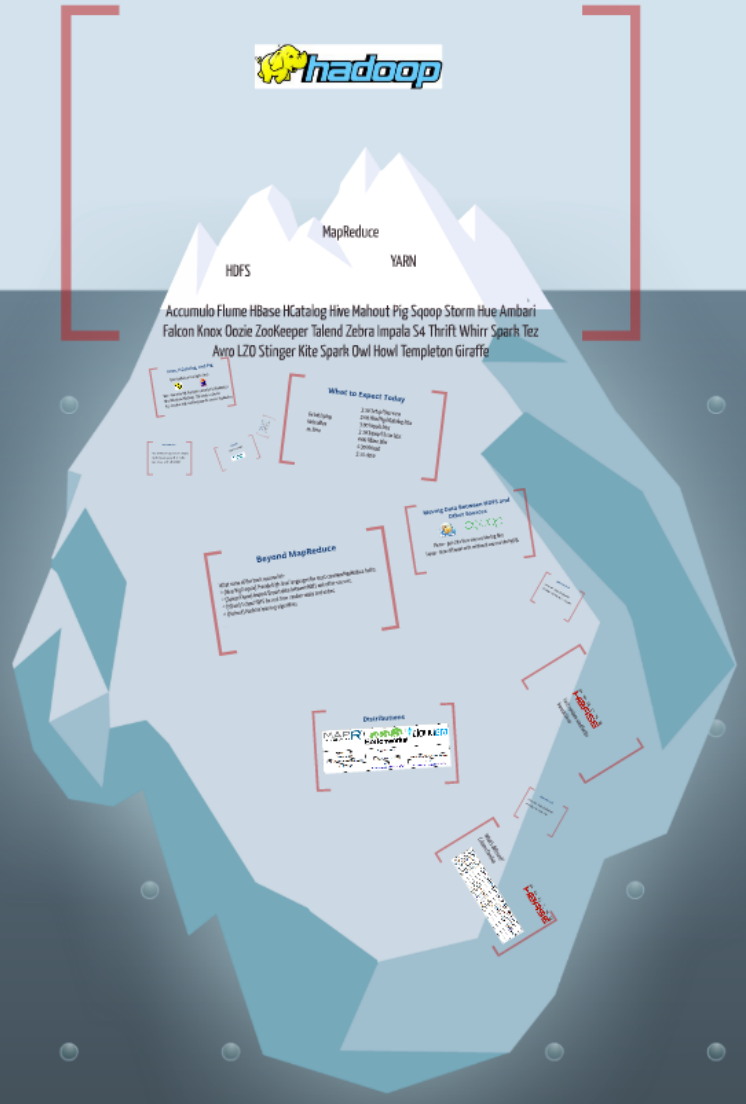
What's different?
Column Families

Row Key	Time Stamp	ColumnFamily `contents`	ColumnFamily `anchor`
"com.cnn.www"	t9		`anchor:cnnsi.com` = "CNN"
"com.cnn.www"	t8		`anchor:my.look.ca` = "CNN.com"
"com.cnn.www"	t6	`contents:html` = "<html>..."	
"com.cnn.www"	t5	`contents:html` = "<html>..."	
"com.cnn.www"	t3	`contents:html` = "<html>..."	

Labs until 4:30

Continue with Cloudera Quickstart VM

Follow HBase Lab in PDF provided



Hadoop's Friends
A Big Data Arsenal

To Start You'll Need:
A 64 bit laptop
Install VirtualBox Now!