Big Data: HDFS Scoop Hive

Ashok

Pathik

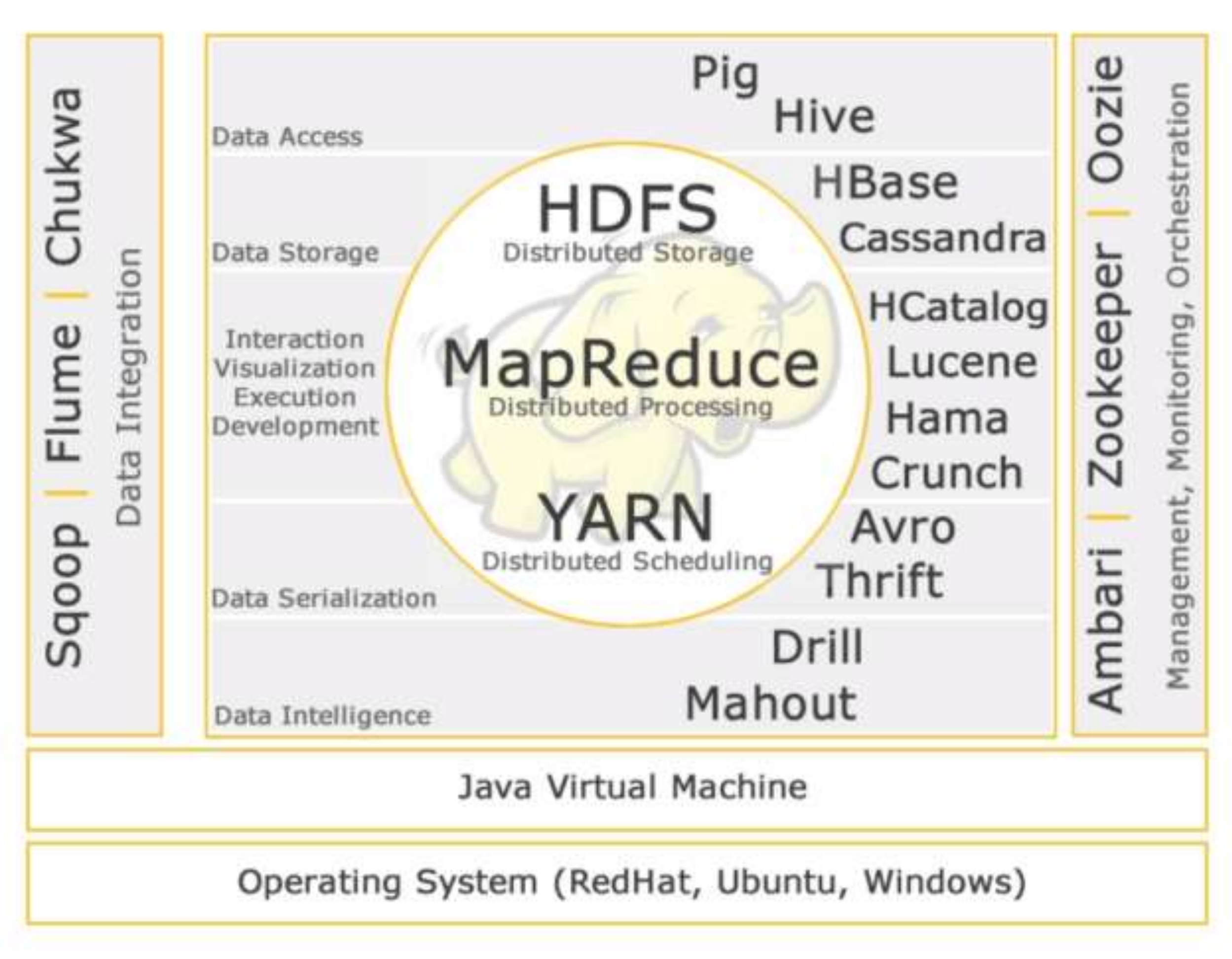
Ravi

Big Data: HDFS Scoop Hive

- 7:00 - 7:15 - Introduction & Recap
- 7:15 - 7:25 - HDFS Commands : Demo
- 7:25 - 7:40 - Word Count with Map Reduce : Demo
- 7:40 - 8:00 - Scoop + Demo (by Ankur Raj)
- 8:00 - 8:05 - Break
- 8:05 - 8:15 - Hive + Demo (by Ravi)
- 8:20 - 8:40 - Flume + Demo
- 8:40 - 9:00 - Impala + Demo (by Ashok)

Hadoop Ecosystem

- Flume
- Sqoop
- Pig
- Hive
- Impala
- Hue



HDFS Commands

- hadoop fs –ls /
- hadoop fs –ls –R /
- hadoop fs –mkdir /input
- hadoop fs –put /home/user1/input/*.xml /input
- hadoop fs –get /input/core-site.xml /home/user1
- hadoop fs –cat /output/1.out
- hadoop fs –tail /output/1.out
- hadoop fs –cp /output/1.out /output2
- hadoop fs –mv /output/1.out /output3
- hadoop job –list
- hadoop fs –rm -r /output
- hadoop jar ….jar wordcount /input /output
- hadoop version

Hadoop Processes(Simiplified)

- Name Node
  - Store HDFS Metadata (which block is on which machine)
- Secondary Name Node
  - Used for housekeeping
- Job Tracker
  - Runs on Master Node
  - Takes Requests and assigns tasks to Task Tracker
- Task Tracker
  - Runs on Data Node
  - Accepts tasks and monitors progress of map reduce tasks
- Data Node
  - Read and Store Data from Disk

Map Reduce : Word Count

- hadoop fs –mkdir /input
- hadoop fs –put /etc/*.conf /input
- hadoop fs -ls -R /input
- hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples-2.6.0-cdh5.5.0.jar
- hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples-2.6.0-cdh5.5.0.jar wordcount /input /output
- hadoop fs –lsr /output
- hadoop tail /output/part-r-00000
- Check the logs by following URL (**SIMILAR** to below)
  - http://quickstart.cloudera:8088/proxy/application_1458576600567_0001/

Sqoop

- Move Data From RDBMS and put it into HDFS
- Move Data From HDFS to RDBMS
- import (to HDFS)
- export (to RDBMS)

- https://sqoop.apache.org/docs/1.4.0-incubating/SqoopUserGuide.html

Hive

- Provides a way for Users Familiar with SQL
  - Leverage experienced “SQL” programmers
- Processes Structured Data
- Data is stored in HDFS
- Metadata/Schema is stored in “metastore”
- Map Reduce Jobs are run to process the data
- Hive Query Language is a subset of “SQL”
- Hive SerDe

Flume

- Store Streaming data in HDFS
  - Example:
    - Collect Web Logs and store in HDFS
  - Terms
    - Source / Sink / Channel / Event
- Ability to handle Multiple sources
- You can chain multiple channels

Impala

- Faster Queries
- User Interface(s)
  - Jdbc / Hue / Impala Shell
- Metadata/Schema is stored in “metastore”
- Similar to Hive Query Language
- Interacts directly with “DataNodes”
- Does not use Map Reduce
- invalidate metadata

- Ref: http://www.cloudera.com/documentation/enterprise/latest/PDF/cloudera-impala.pdf