# GCD Theory week 4 – Hadoop / Hive

**Objectives**

The student:

* Understands the concept behind HDFS (NameNodes, redundancy)
* Is able to store and retrieve files in Hadoop
* Is able to use Hive as a data warehouse for storing data
* Is able to make simple queries on the Hive database

**Preparation**

Watch the following videos from Udacity Intro to Hadoop and MapReduce. The code ‘Lesson 1.13’ means within ‘Lesson 1’, the 13rd part of the lesson. The MapReduce technology will be discussed in a later lesson.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Title** | **Duration** | **Udacity Lesson** | **Week** | **Remarks** |
| Doug Intro | 0:24 | 1.13 | 4 |  |
| Doug Cutting: The Origins of Hadoop | 3:11 | 1.14 | 4 |  |
| Hadoop Logo Intro | 0:20 | 1.15 | 4 |  |
| Doug Cutting: The Name of Hadoop | 1:01 | 1.16 | 4 |  |
| Core Hadoop | 0:47 | 1.17 | 4 |  |
| Hadoop Ecosystem | 3:32 | 1.18 | 4 |  |
| HDFS | 1:50 | 2.1 | 4 |  |
| Data Redundancy | 1:16 | 2.2 | 4 |  |
| NameNode Standby | 1:04 | 2.3 | 4 |  |
| HDFS Demo | 2:17 | 2.4 | 4 | Open your CDH Virtual Machine and type all statements made in the video on your Virtual Machine as well! |
| HDFS | quiz | Pr 2.1 | 4 |  |
| DataNode | quiz | Pr 2.2 | 4 |  |
| NameNode | quiz | Pr 2.3 | 4 |  |

For Hive introduction, watch the following Introduction course (14:49): <http://www.cloudera.com/content/cloudera/en/resources/library/training/introduction-to-apache-hive.html>

Remark: try not to run the ‘Shakespeare’ example in your CDH Virtual Machine. This will not work, due to an older version of CDH used in the presentation.