Lesson 1: Introduction to Apache Spark

- **Video:** Introduction to Apache Spark 8 min
- **Video:** Architecture of Spark 7 min
- **Reading:** Setup PySpark on the Cloudera VM 10 min
- **Reading:** Lesson 1: Intro to Apache Spark - Slides 10 min
- **Quiz:** Spark Lesson 1 6 questions

Lesson 2: Resilient Distributed Datasets and Transformations

- **Video:** Resilient Distributed Datasets 10 min
- **Video:** Spark Transformations 10 min
- **Video:** Wide Transformations 10 min
- **Reading:** Lesson 2: RDD and Transformations - Slides 10 min
- **Quiz:** Spark Lesson 2 5 questions
- **Programming Assignment:** Simple Join in Spark 3h

Lesson 3: Job scheduling, Actions, Caching and Shared Variables

- **Video:** Directed Acyclic Graph (DAG) Scheduler

[One time setup] Install IPython

From the top left menu, Open a terminal:
Applications => System Tools => Terminal

Type:

```
sudo easy_install ipython==1.2.1
```

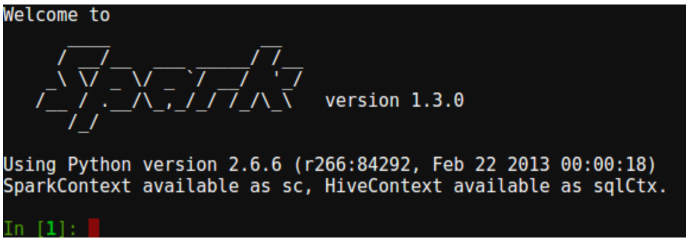
Hit enter, administrator password is **cloudera**.

Launch pyspark with IPython

Every time you need to open the pyspark shell, open a terminal and type:

```
PYSPARK_DRIVER_PYTHON=ipython pyspark
```

Hit enter, after the startup logs, you should see the pyspark console:



```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 1.3.0
      /_/

Using Python version 2.6.6 (r266:84292, Feb 22 2013 00:00:18)
SparkContext available as sc, HiveContext available as sqlCtx.

In [1]:
```

Check version

To make sure that PySpark started correctly, print out the version by typing in the PySpark IPython terminal: