К счастью, данные уже представленны в простом CSV-формате, и не требуют упрошения или другой подготовке для использования с SparkMlib. В последствии было бы интересно исследовать(изучить) некоторые преобразования данных, но для начала их можно использовать как есть. Файл covtype.data должен быть извлечен(из архива/разархивирован) и скопирован в HDFS. Эта глава предполагает, что файл доступен в /user/ds/-директории.
Запустите оболочку Spark (Spark-shell)
Абстракция Spark Mlib для вектора свойств (вектор-функции) известна как LabeledPoint, котороя состоит из вектор-функций Spark Mlib и целевого значения, называемого здесь меткой. Цель(целевое значение) это переменная размера Double, а Вектор, собственно, абстракция поверх нескольких Double-переменных. Это озачает(предполагает), что LabeledPpoint предназначен для использования исключительно с числовыми функциями. Он может использоваться с категориальными функциями с подходящей кодировкой.
Одной из таких кодировок является one-hot или 1-of-n кодировка, при которой одна категориальная функция, которая может принимать N значений, представляется в виде N числовых функций, каждая из которых может принимать значение 0 или 1. Исключительно одна из этих функций может равняться 1, в то время как другие должны равняться 0. Например категгориальная функция погоды, которая может принимать значения “пасмурно”, “дождь” или “ясно” будет представлена в виде трех числовых функций, где “пасмурно” представлено в виде 1,0,0; “дождь” - 0,1,0; и т.д. Эти три числовые функции можно рассматривать как is_cloudy, is_rainy и is_clear. Другая возможная кодировка просто-напросто присваивает определенное числовое значение каждому возможному значению категориальной функции. Например, “пасмурно” будет 1.0, “дождь” - 2.0 и т.д.