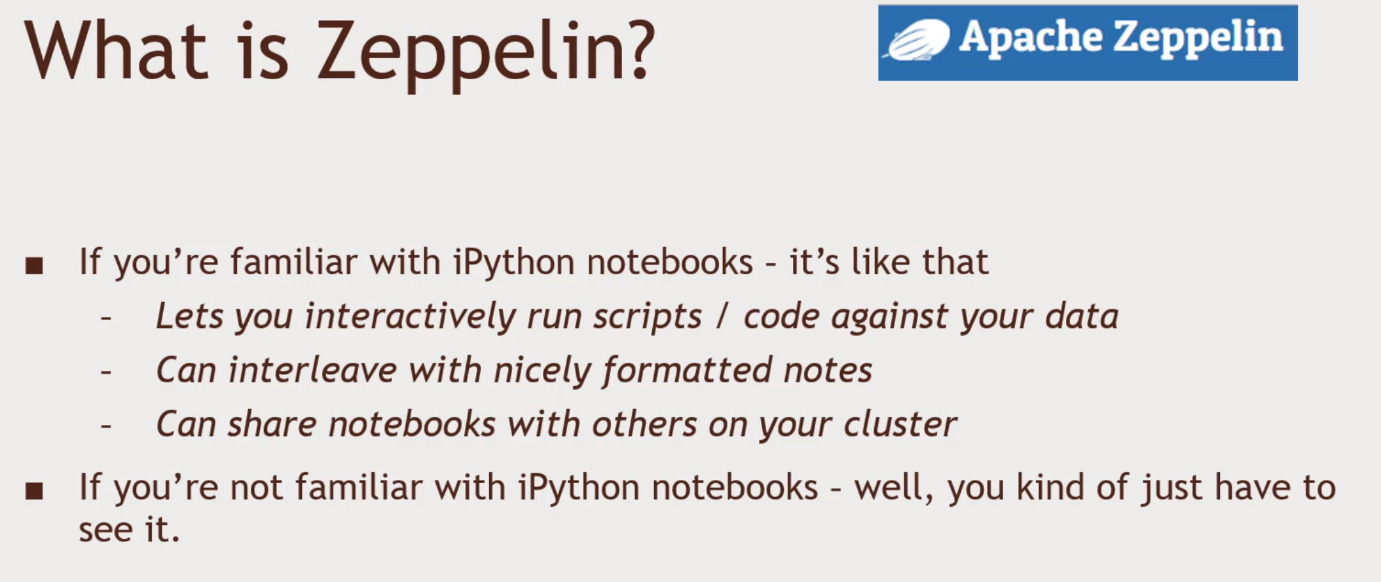
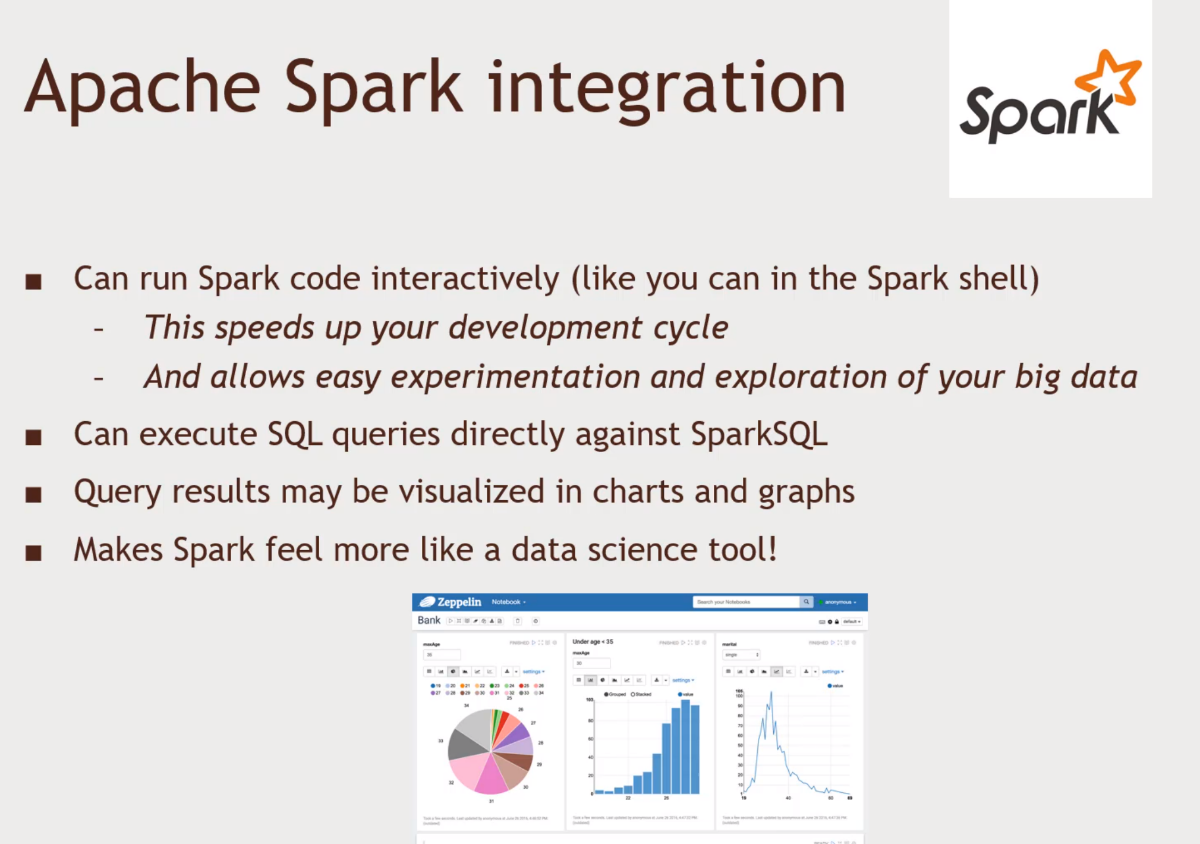
Apache Zappelin – Quick way to experiment with your Spark Script.

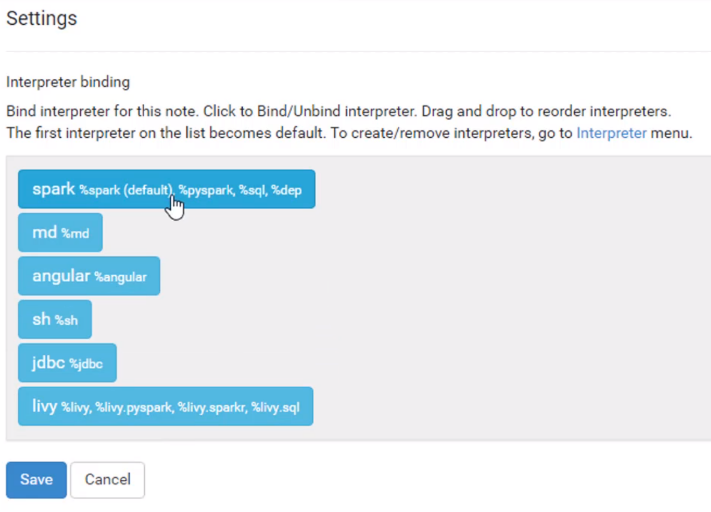
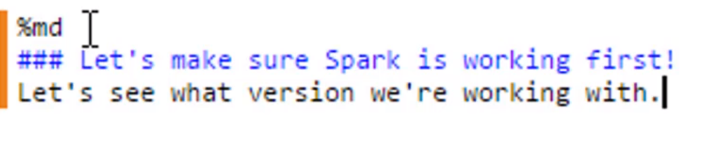
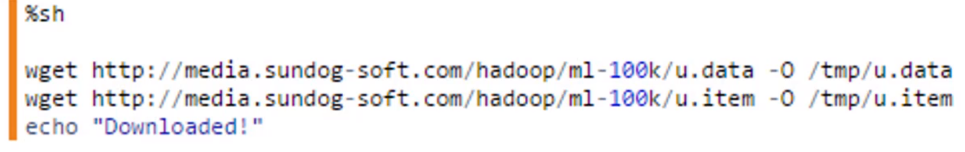
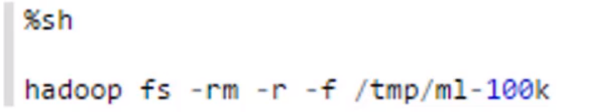
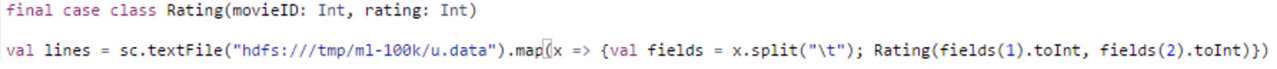
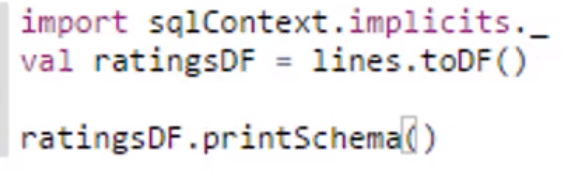
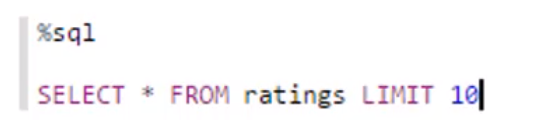
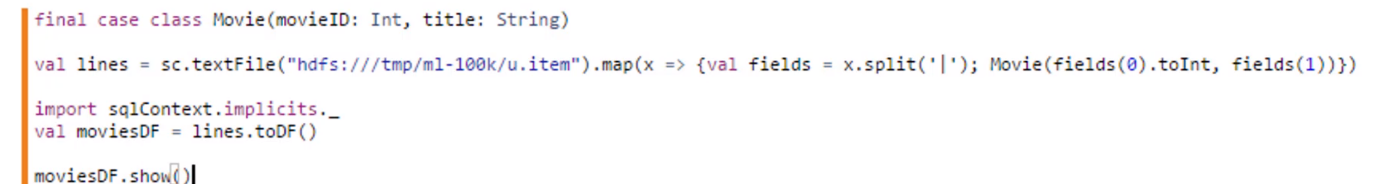
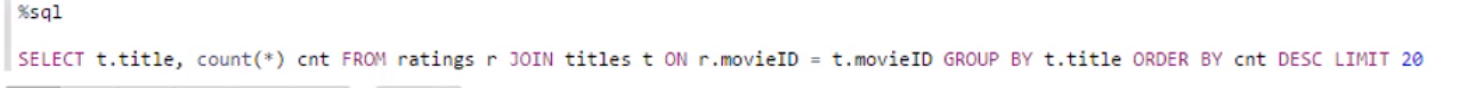
Let you experiments and etc.

Notebook Interface to your big data





Hortonworks preinstall zeppelin

1. Connect to zeppelin 127.0.0.1:9995
2. Notebook interface to your cluster
   1. Click on the create new note button and name it
   2. Run spark code
   3. Click on the gear icon ‘Interpreter binding’
   4. Type in these’ and + Shift Enter to see if spark is working
   5. Remember: we are using spark by default as our primary backend
      1. sc.version. spark 1.6.2
      2. using scala programming language
      3. can execute shell commands by
      4. Shell command to upload it into HDFS so that you can run in your cluster checking if the directory exists
      5. If not type in ‘hadoop fs -mkdir /tmp/ml-100k’
      6. Upload our data ‘hadoop fs -put /tmp/u.data /tmp/ml-100k’ and ‘hadoop fs -put /tmp/u.item /tmp/ml-100k’
      7. Hit the gear icon to show title. Its like jupyter notebook.
   6. Write Scala code
      1. import the files as a RDD
      2. Create the dataframe from the RDD
      3. Query to see the top movies
3. Expose the info into SQL form
   1. Code the following in Zeppelin and expose it
      1. Create the table
   2. Query using SQL
      1. ‘SELECT rating, COUNT(\*) as count FROM ratings GROUP BY rating’
   3. You can do visualisation too.
   4. Create another dataframe. By following the above steps
   5. Expose the moviesDF to the HDFS
      1. moviesDF.registerTempTable(‘titles’)
   6. Do SQL Queries