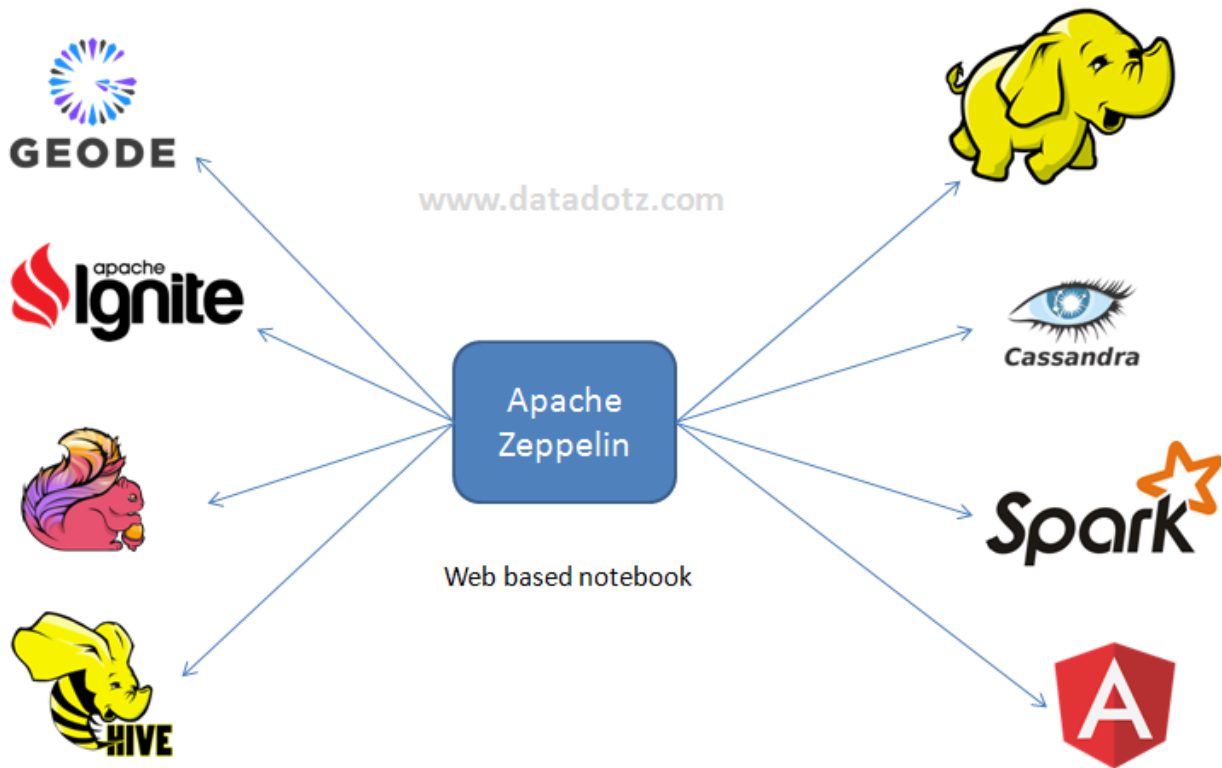


Apache Zeppelin: Quick start with Apache Spark

Apache Zeppelin web-based notebook that enables interactive data analytics against multiple language basket. It also provides apache spark in additional to data visualization. Currently **Apache Zeppelin is in incubation**

for more please refer : <https://zeppelin.incubator.apache.org/>

Current version zeppelin is : Zeppelin-0.5.5



Currently this tutorial is for Zeppelin with Apache Spark

Perquisite

1. Git
2. Npm (NPM is NodeJs package manager)
3. Java 1.7
4. Apache spark for this quick start

Installation Perquisite

This tutorial is written using ubuntu os. For other Linux os please refer similar commands or write to us

sudo apt-get update

sudo apt-get install git

sudo apt-get install npm (NPM is NodeJs package manager)

Download the Java from Oracle. Please Check for Oracle Website if the link is broken. Please check for current version in the Oracle

Download jdk 1.7 or higher because java is the dependency for zeppelin.

<http://www.oracle.com/technetwork/java/javase/downloads/jdk7-downloads-1880260.html>

Set the Environment path of java in .bashrc

export JAVA_HOME=/home/datadotz/jdk1.7.0_45

Note : Even you can set this in .bash_profile also

Download Zeppelin

Download the Zeppelin from Zeppelin site. Please Check for Zeppelin Website if the link is broken. Please check for current version in the Zeppelin

<http://www.us.apache.org/dist/incubator/zeppelin/0.5.5-incubating/zeppelin-0.5.5-incubating-bin-all.tgz>

Set Environment path and variable in Zeppelin

con/zeppelin-env.sh

export JAVA_HOME=/home/datadotz/jdk1.7.0_45

export SPARK_HOME=/home/datadotz/spark-1.5.1-bin-hadoop2.6

Note : you can also use the latest version of spark -1.5.2

Zeppelin Installation

Please use the bellow command to start

zeppelin-daemon.sh start

This will start Zeppelin server daemon can check the status by using command jps (java process status)

Zeppelin UI

Please check your web page by default Zeppelin UI runs on port 8080

localhost:8080

#--- Want to change the UI port number change the port number in zeppelin-site.xml --#

Spark installation

Please Refer to chennaihug.org for spark installation

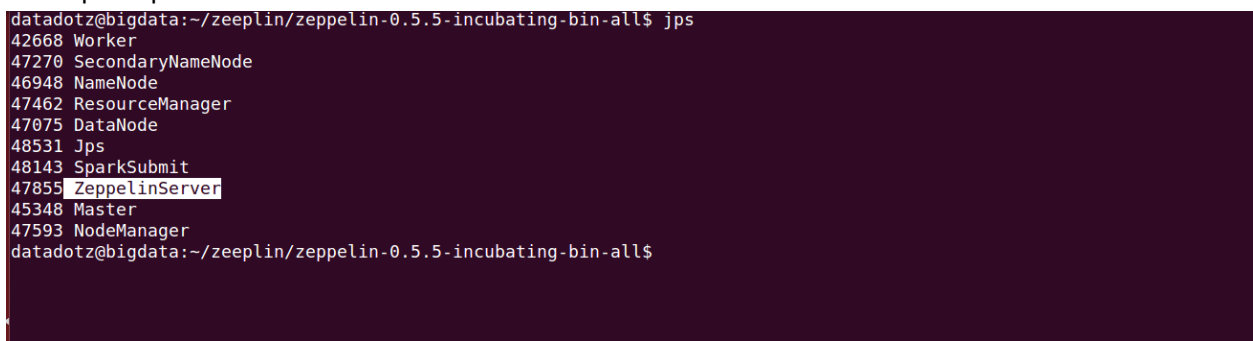
<http://chennaihug.org/knowledgebase/spark-master-and-slaves-single-node-installation/>

Command to start

sbin/start-all.sh

List of all Daemons Fig1

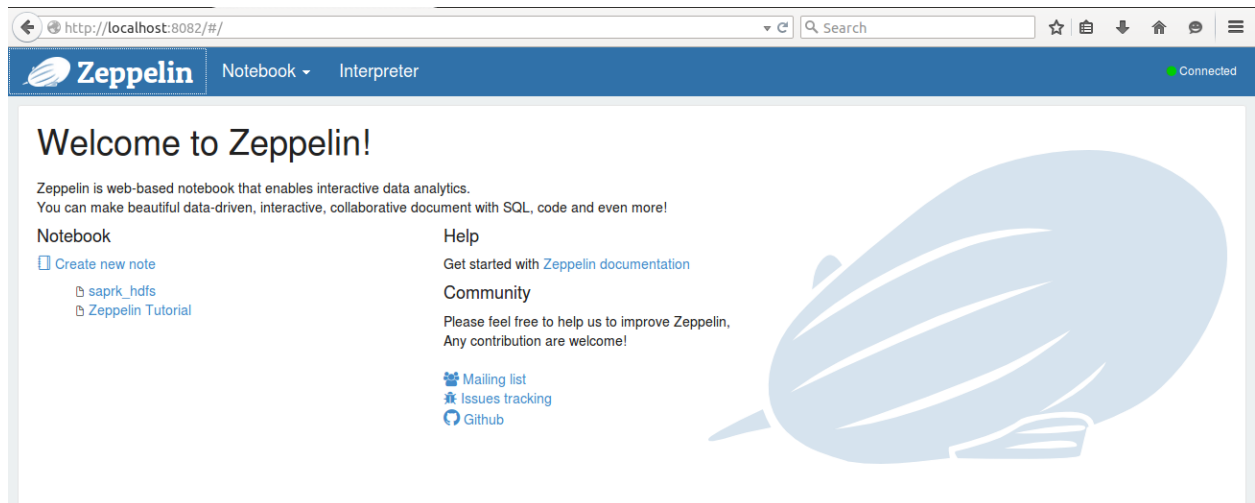
This figure shows the list of all daemons running . If just using zeppelin with spark then no need of Hadoop to up and run.



```
datadotz@bigdata:~/zeplin/zeppelin-0.5.5-incubating-bin-all$ jps
42668 Worker
47270 SecondaryNameNode
46948 NameNode
47462 ResourceManager
47075 DataNode
48531 Jps
48143 SparkSubmit
47855 ZeppelinServer
45348 Master
47593 NodeManager
datadotz@bigdata:~/zeplin/zeppelin-0.5.5-incubating-bin-all$
```

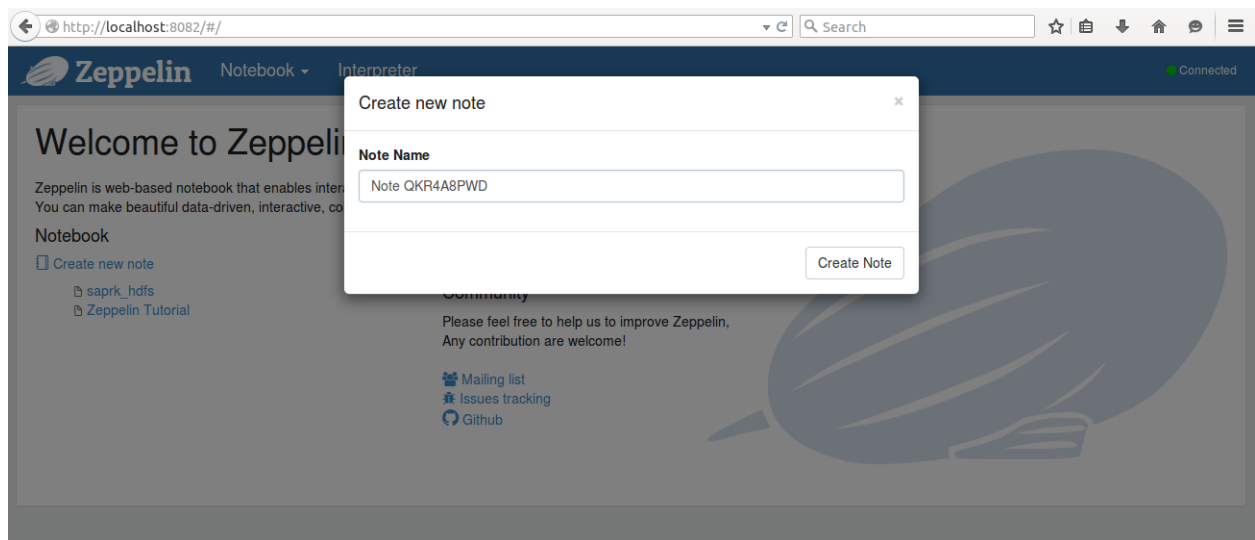
1.Web UI Fig2

This figure shows the default web UI of Zeppelin which runs on IP address localhost and the port on 8082(in my case I changed the port number from 8080 to 8082)



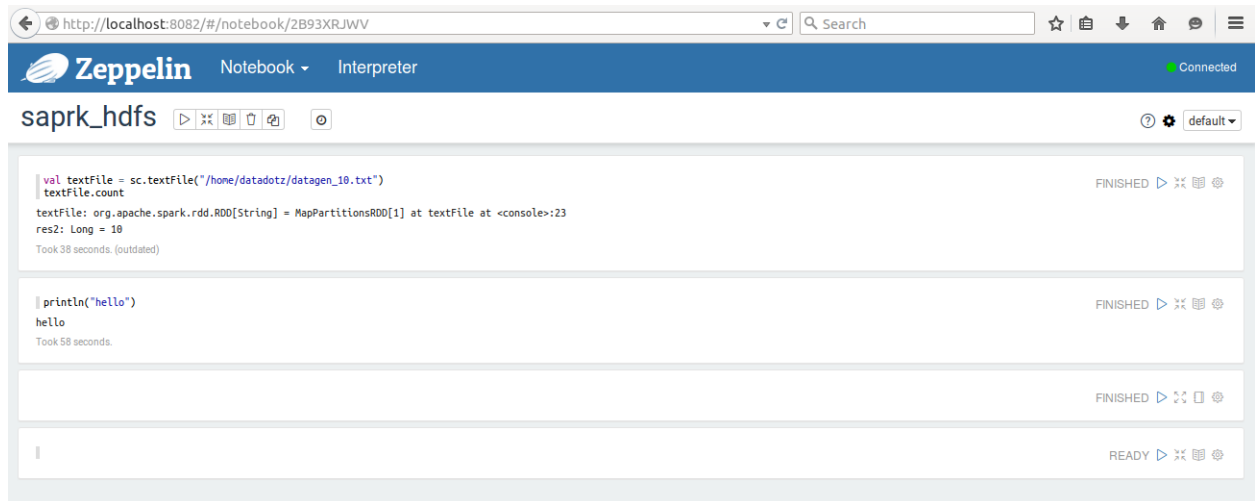
2. Note book Fig3

This is how you have to create a note book . Simply to say note book is like a editor where you can run the commands and scripts



3. Use apache spark Interpreter Fig 4

This just a sample example to load a file form my Linux in to spark . And testing it by running count



Analysis of drug data using apache spark SQL and Zeppelin

Load the data and create a schema and temporary table

Table name : customers

Input Data set : datagen_10.txt (drug data set)

```
1,Brandon Buckner,avil,female,525
2,Veda Hopkins,avil,male,633
3,Zia Underwood,paracetamol,male,980
4,Austin Mayer,paracetamol,female,338
5,Mara Higgins,avil,female,153
6,Sybill Crosby,avil,male,193
7,Tyler Rosales,paracetamol,male,778
8,Ivan Hale,avil,female,454
9,Alika Gilmore,paracetamol,female,833
10,Len Burgess,metacin,male,325]
```

Query : To find the total amount of the drugs

"select drug , sum(amt) from customers group by drug"

The output of the query is to return all the records based of sum of amount group by drug

Output in Various forms

Form 1 Fig 5

This web UI shows the tabular view of output

The screenshot shows the Zeppelin Notebook interface. The top bar includes the Zeppelin logo, 'Notebook' dropdown, 'Interpreter' dropdown, and a 'Connected' status indicator. The notebook name is 'saprk_hdfs'. The main area displays a Scala query that reads a text file, defines a 'Customer' case class, and registers a temporary table named 'customers'. Below the Scala code, the output is shown as a tabular view. The table has two columns: 'drug' and '_c1'. The data rows are: paracetamol (2,929), metacin (325), and avil (1,958). The status 'FINISHED' is shown next to the code and the table.

```
val x = sc.textFile("/home/datadotz/datagen_10.txt")
case class Customer(sno: Int, name: String, drug: String, gender: String, amt: Int)
val dfCustomers = x.map(_split(",")).map(p => Customer(p(0).trim.toInt, p(1), p(2), p(3), p(4).trim.toInt)).toDF()
dfCustomers.registerTempTable("customers")

x: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[38] at textFile at <console>:23
defined class Customer
dfCustomers: org.apache.spark.sql.DataFrame = [sno: int, name: string, drug: string, gender: string, amt: int]

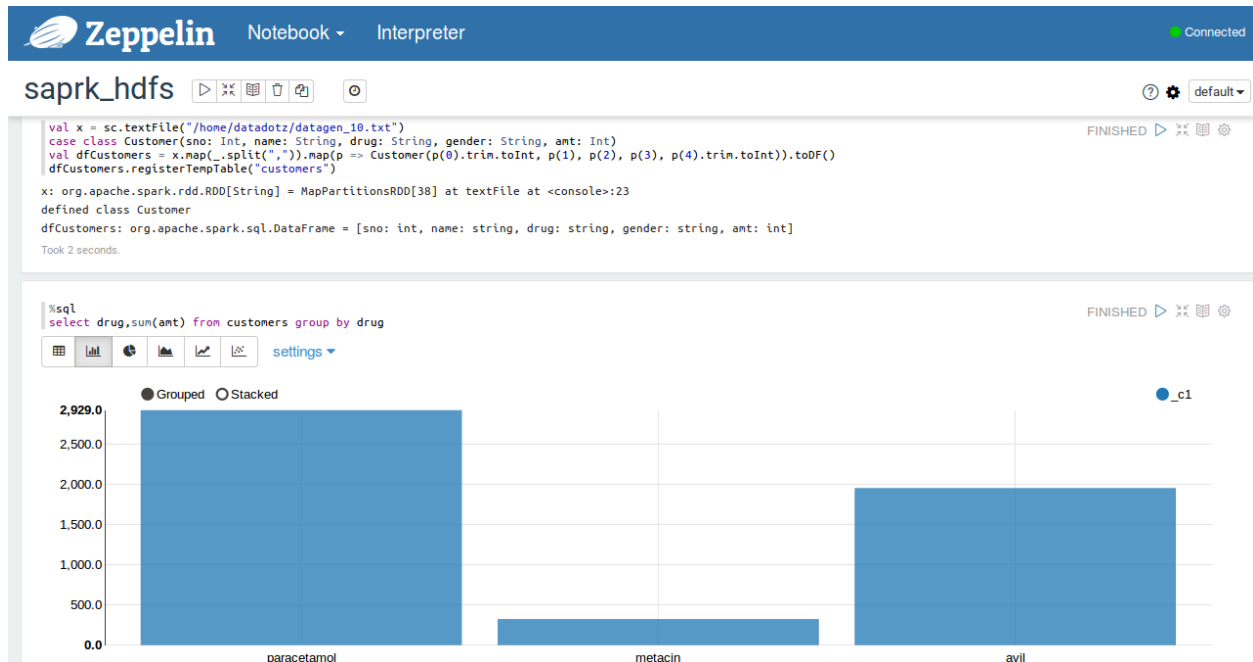
Took 2 seconds.
```

```
%sql
select drug,sum(amt) from customers group by drug
```

drug	_c1
paracetamol	2,929
metacin	325
avil	1,958

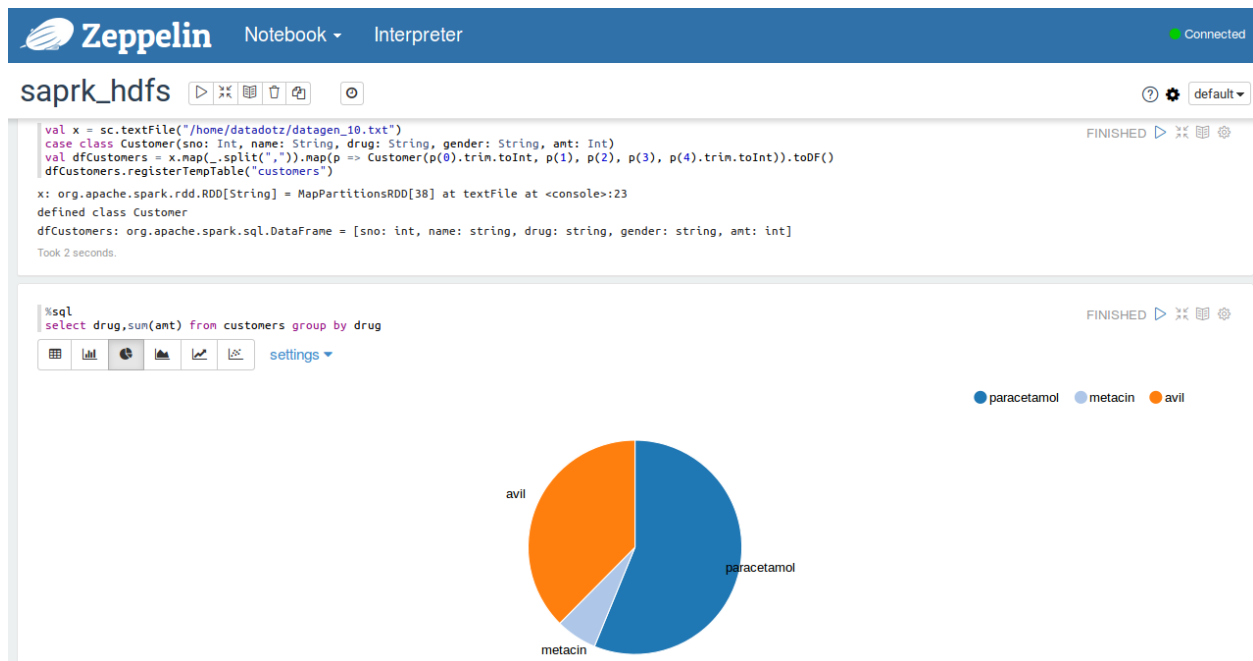
Form 2 Fig 6

This web UI shows the bar chart representation

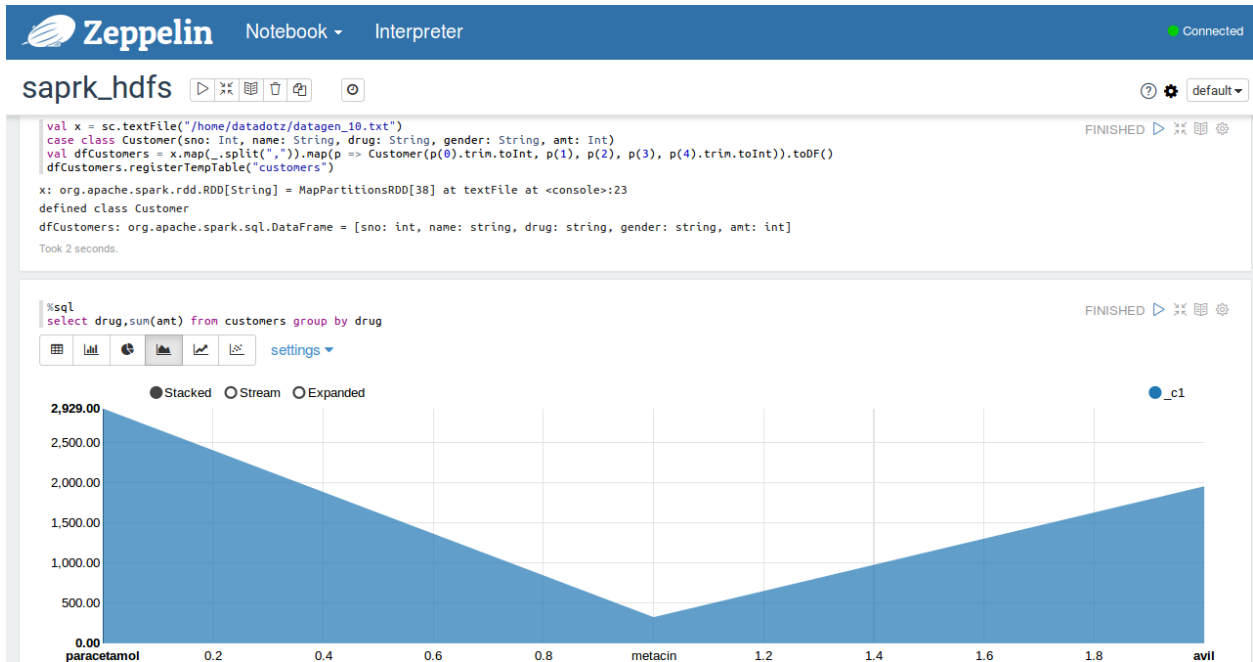


Form 3 Fig 7

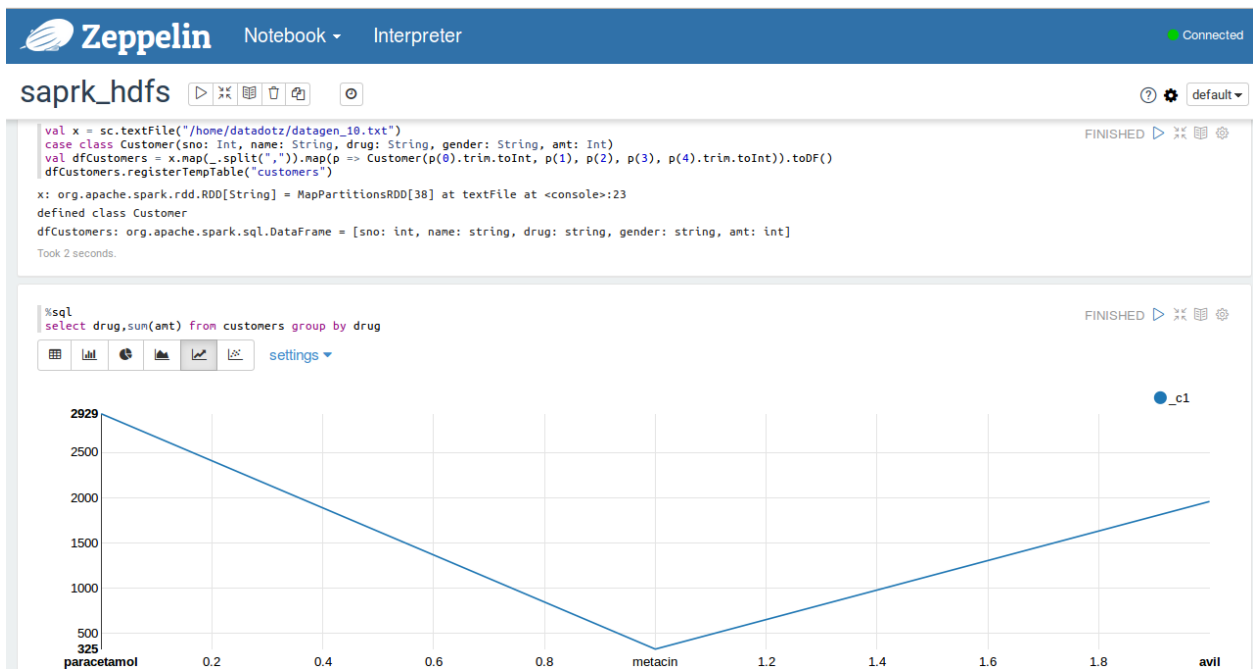
This web UI shows the pie chart representation



Form 4 Fig 8



Form 5 Fig 9



Form 6 Fig 10

