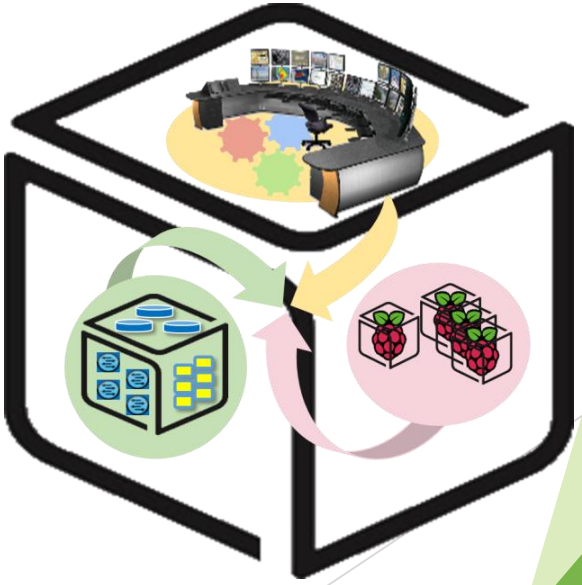
SmartX Labs
for Computer Systems

Analytics Lab
(2016, Spring)



NetCS Lab

History and Contributor of Cluster Lab (2016. 05. 02.)

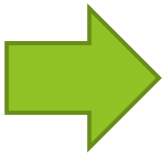
Version	Updated Date	Updated Contents	Contributor
v2r2	2015/10	(구) Analytics Lab 작성	송지원
v3	2016/05	Analytics Lab 수정	송지원

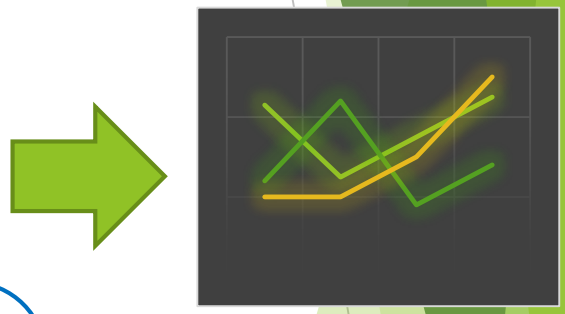
CSLab: Analytics LAB - Goal

- Data Processing with Spark & Zeppelin

Data Processing & Visualization

Data



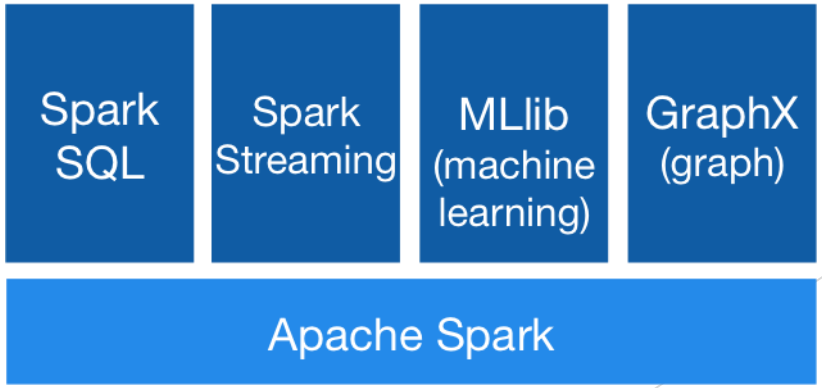


Apache Spark
- Concept

Apache Spark™ is a fast and general engine for large-scale data processing.

- In-memory data processing framework: Fast!
- Easy to use, community fastly growing
- Libraries: SQL and DataFrame, Streaming, MLlib, GraphX
- Run on standalone or Mesos, Yarn, etc

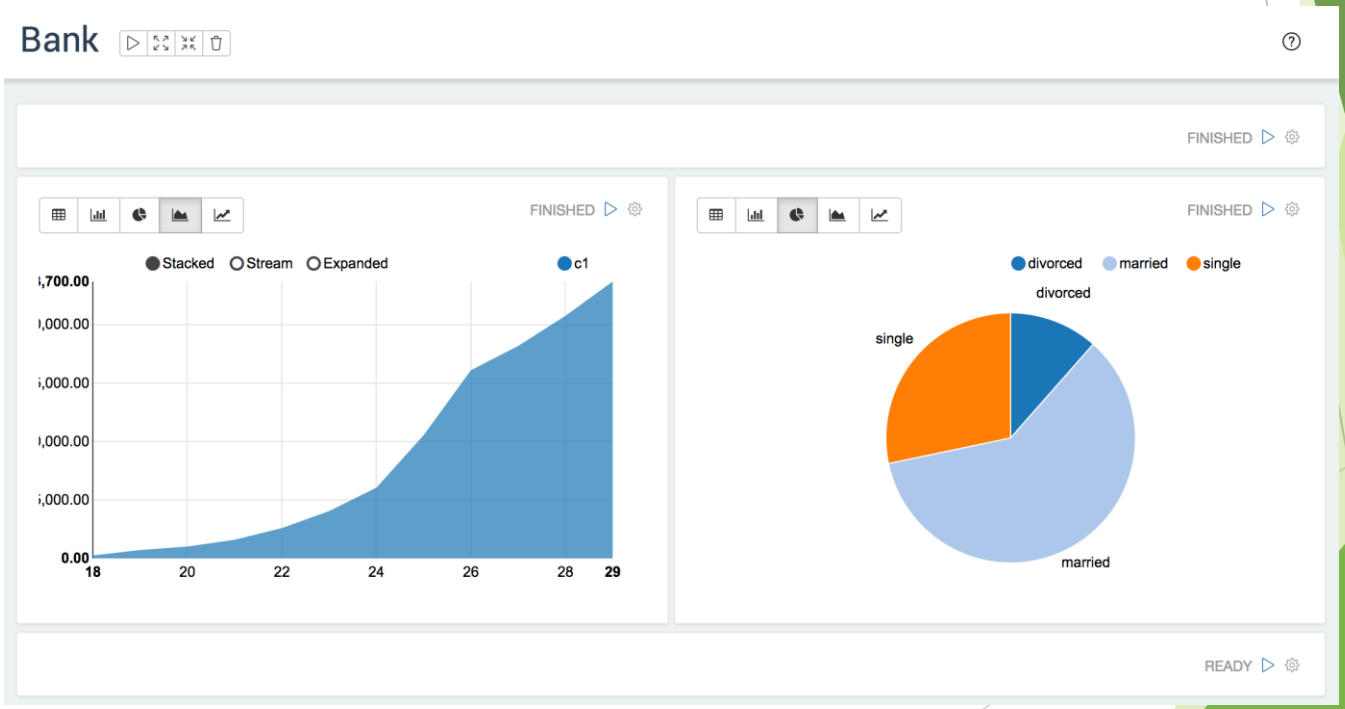
- Scala, Java, Python



Apache Zeppelin
-Concept

A web-based notebook that enables interactive data analytics.

Support Spark



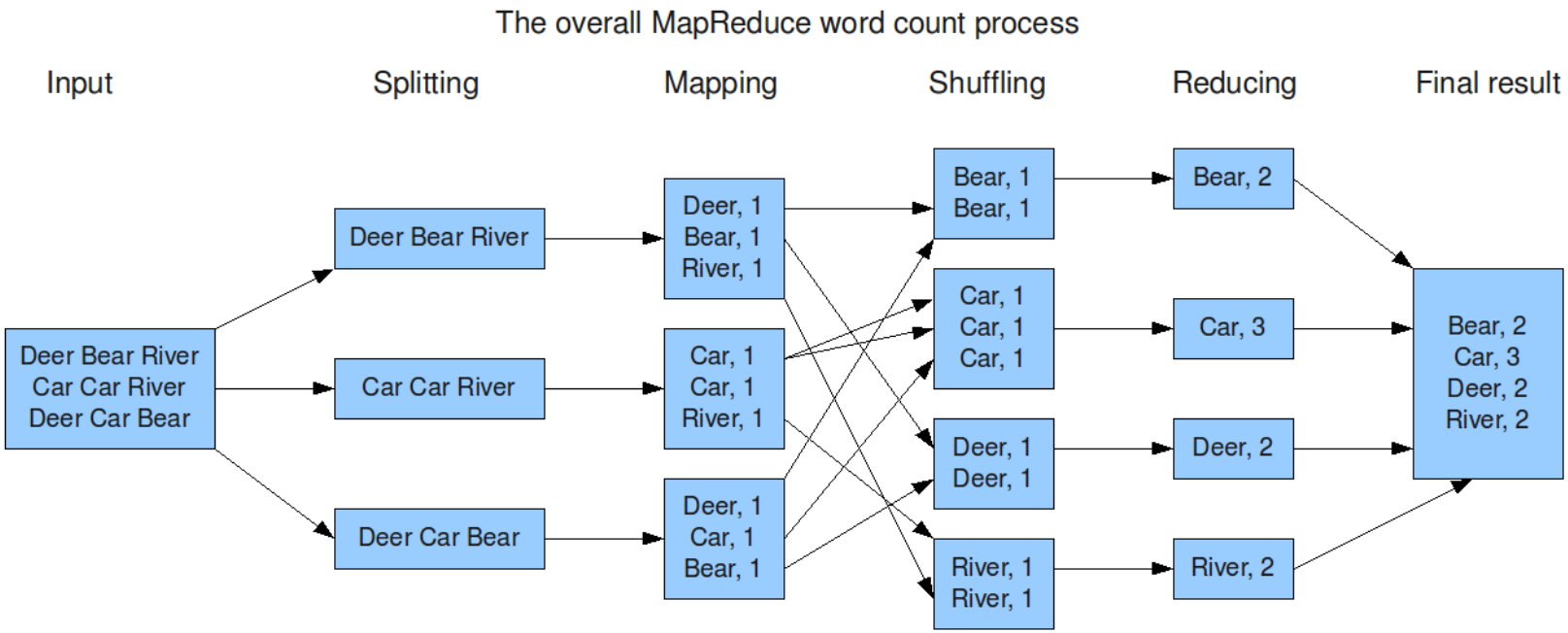
Tip. Pyspark: Spark in Python - Sample

Zeppelin tutorial converted to pyspark

```
%pyspark

from pyspark.sql.types import *

zeppelinHome = os.getcwd()
bankText = sc.textFile(zeppelinHome + "/data/bank.csv")

bankSchema = StructType([StructField("age", IntegerType(), False),
            StructField("job", StringType(), False),
            StructField("marital", StringType(), False),
            StructField("education", StringType(), False),
            StructField("balance", IntegerType(), False)])

bank = bankText.map(lambda s: s.split(";")).filter(lambda s: s[0] != "\"age\"").map(lambda s: (
            int(s[0]),
            str(s[1]).replace("\"", ""),
            str(s[2]).replace("\"", ""),
            str(s[3]).replace("\"", ""),
            int(s[5])))

bankdf = sqlContext.createDataFrame(bank, bankSchema)
bankdf.registerTempTable("bank")
```

```
# In zeppelin directory, make data directory and download sample data file.
$ cd zeppelin-0.5.5-incubating-bin-all
$ mkdir data
$ cd data
$ wget https://s3.amazonaws.com/apache-zeppelin/tutorial/bank/bank.csv
```

2. Processing Big Data
- Map and Reduce



2. Processing Big Data
- Map and Reduce in Spark

RDD (Resilient Distributed Datasets): a distributed memory abstraction that allows programmers to perform in-memory computations on large clusters while retaining the fault tolerance of data flow models like MapReduce.

class pyspark.RDD

map()
groupBy(), groupByKey()
reduce(), reduceByKey()
join()
sort(), sortByKey()
union()
…

http://spark.apache.org/docs/latest/api/python/pyspark.html

3. Apache Zeppelin
- Wordcount

```
# Prepare data
cd zeppelin-0.5.5-incubating-bin-all
mkdir data
cd data
wget https://www.dropbox.com/s/dvtrxdr8am49yvv/wordcount.txt
Hadoop fs –put wordcount.txt /
```

wordcount.txt: **Remarks by President Obama at Hankuk University (**https://www.whitehouse.gov/the-press-office/2012/03/26/remarks-president-obama-hankuk-university**)**

3. Apache Zeppelin
- Wordcount

```
%pyspark

from pyspark.sql.types import *
import os

zeppelinHome = os.getcwd()
lines = sc.textFile(zeppelinHome + "/data/wordcount.txt")
counts = lines.flatMap(lambda x: x.split(' ')) \
        .map(lambda x: (x, 1)) \
        .reduceByKey(lambda a, b: a + b)

countSchema = StructType([
    StructField("word", StringType(), True),
    StructField("counts", IntegerType(), True)])
countdf = sqlContext.createDataFrame(counts, countSchema)
countdf.registerTempTable("wordcount")
```

(If you have trouble using Zeppelin,
restart Zeppelin daemon.)
$ bin/zeppelin-daemon.sh restart

3. Apache Zeppelin
- Wordcount (Result)

```
%sql
select word, counts value
from wordcount
where counts > 20
order by counts
```

FINISHED

word	value
for	21
with	22
have	24
this	26
will	28
And	38
I	40
is	42
our	44
we	47

Took 0 seconds.

3. Apache Zeppelin
- Data Cleaning

Actually, there are 26 ‘have’, 37 ‘this’ in wordcount.txt

We need to remove punctuation marks like ‘.’, ‘,’, etc.

It can be done by re.sub() and map() functions.

Add some codes to make dirty data clean and get the right answer.

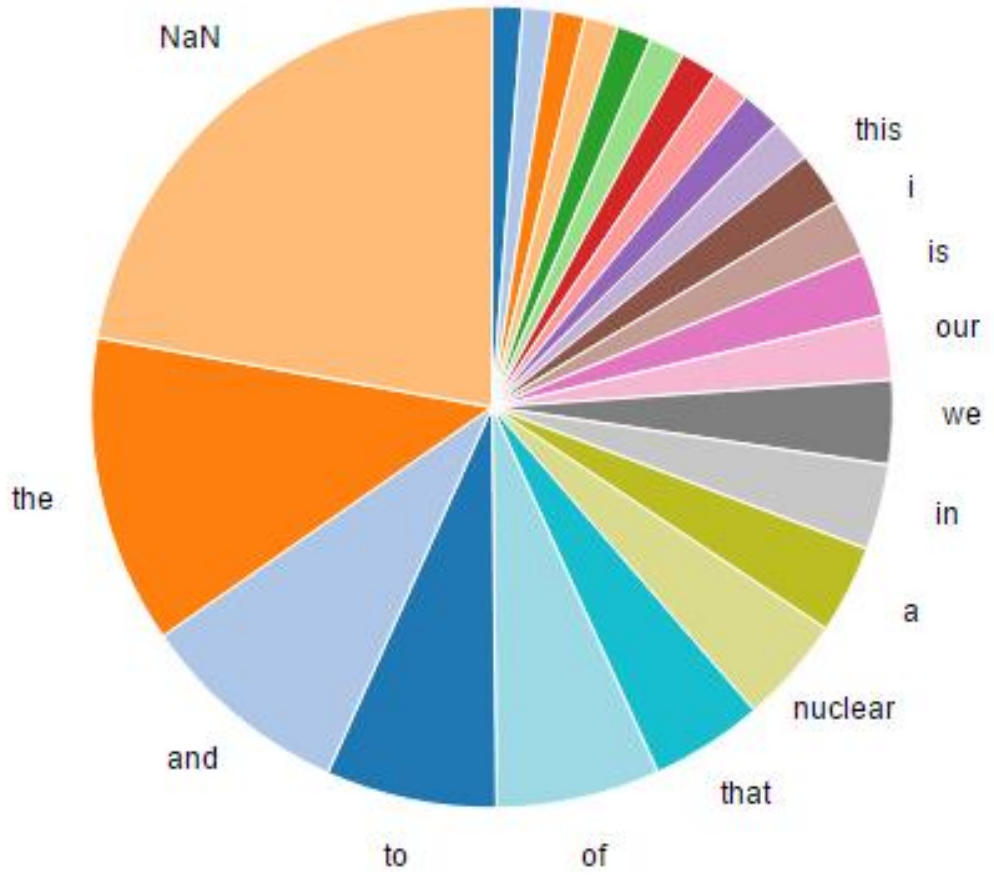
This is to make complex model simple, or get better answer.

3. Apache Zeppelin
- Wordcount (fixed)

word	value
world	21
as	21
but	22
with	23
you	24
weapons	24
have	26
for	26
it	28
will	29

Took 0 seconds.

3. Apache Zeppelin
- Wordcount - graph



3. Apache Zeppelin
- Wordcount - graph

```
%sql
select word, counts value
from wordcount
where counts > 20
```

FINISHED

settings

All fields:

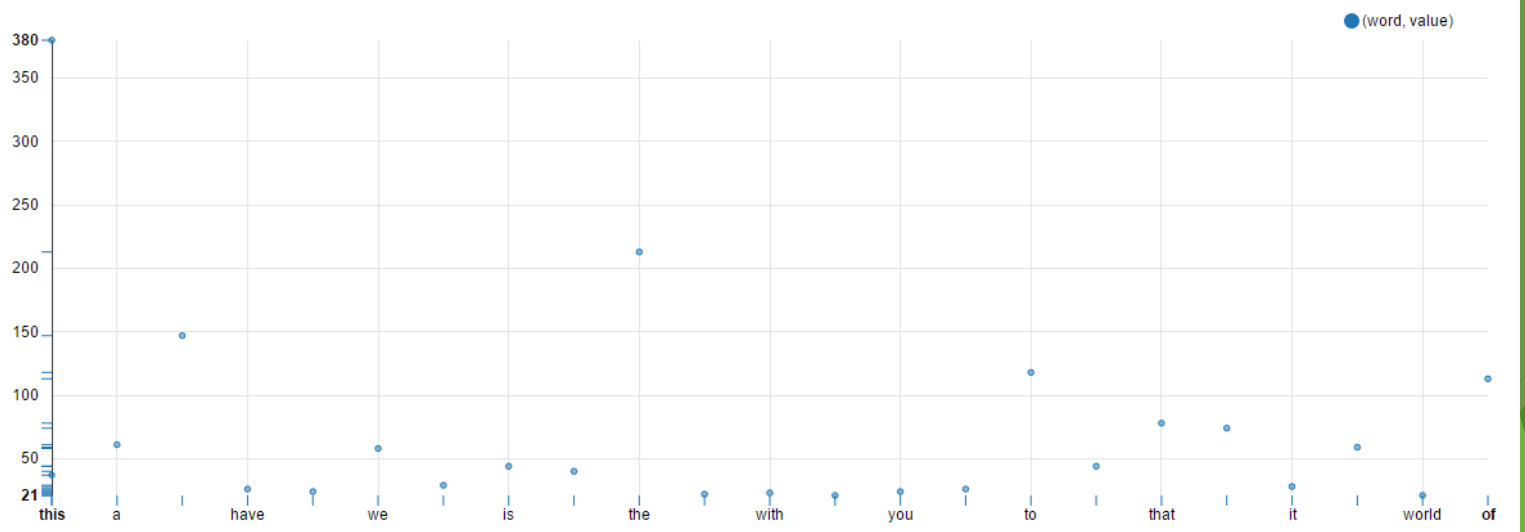
word value

xAxis: word

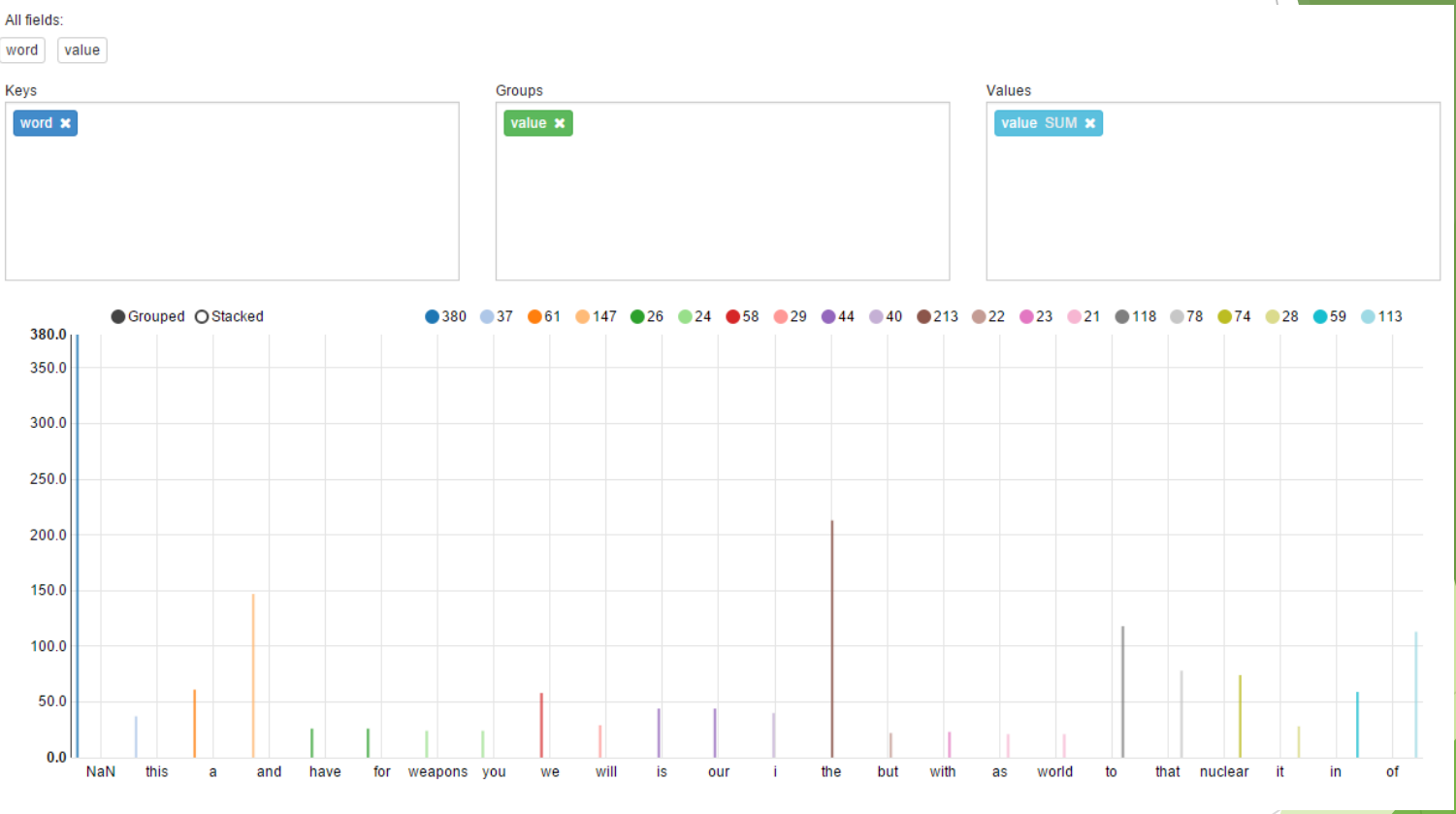
yAxis: value

group

size



3. Apache Zeppelin
- Wordcount - graph



Thank You for Your Attention Any Questions?