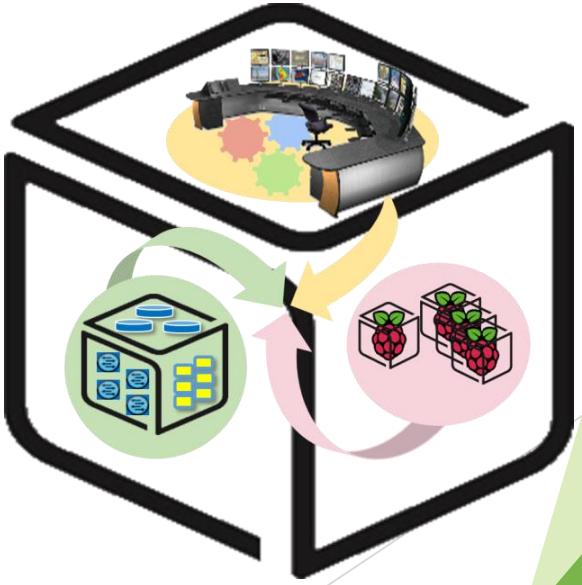
SmartX Labs
for Computer Systems

Analytics Lab
(2016, Spring)



NetCS Lab

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

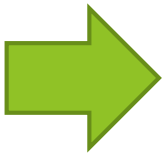
Version	Updated Date	Updated Contents	Contributor
v2r2	2015/10	(구) Analytics Lab 작성	송지원
v3	2016/05	Analytics Lab 수정	송지원
v4	2016/06/07	검수자 피드백 반영 및 내용 수정	송지원
v4r1	2016/06/08	Wordcount 예제 설명 추가	송지원

CSLab: Analytics LAB - Goal

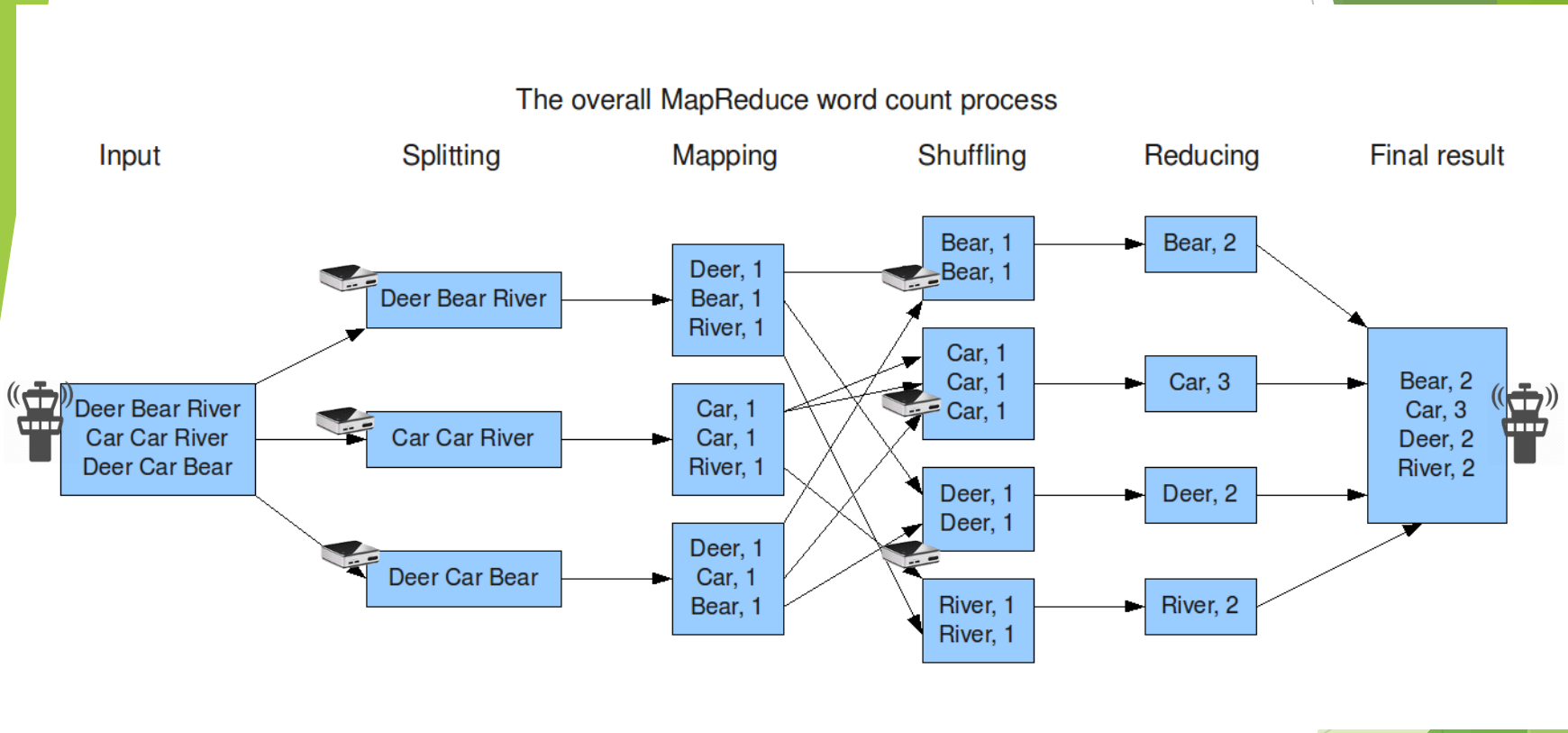
- Data Processing with Spark & Zeppelin

Data Processing & Visualization

Data



1. Background
- Processing Big Data: MapReduce

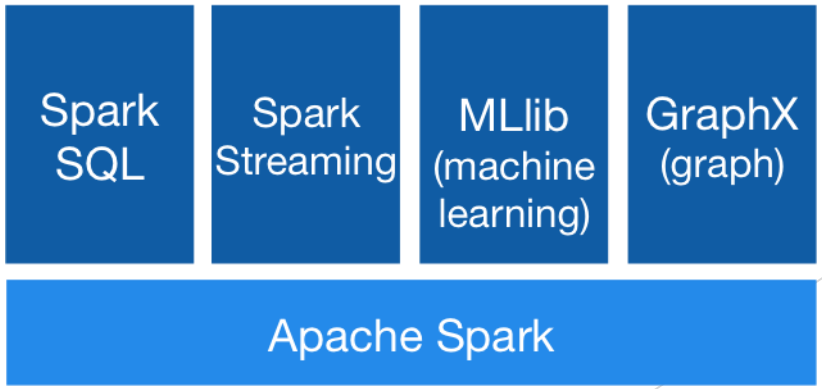


1. Background
- Apache Spark

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



2. Background
- Spark RDD and APIs

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.

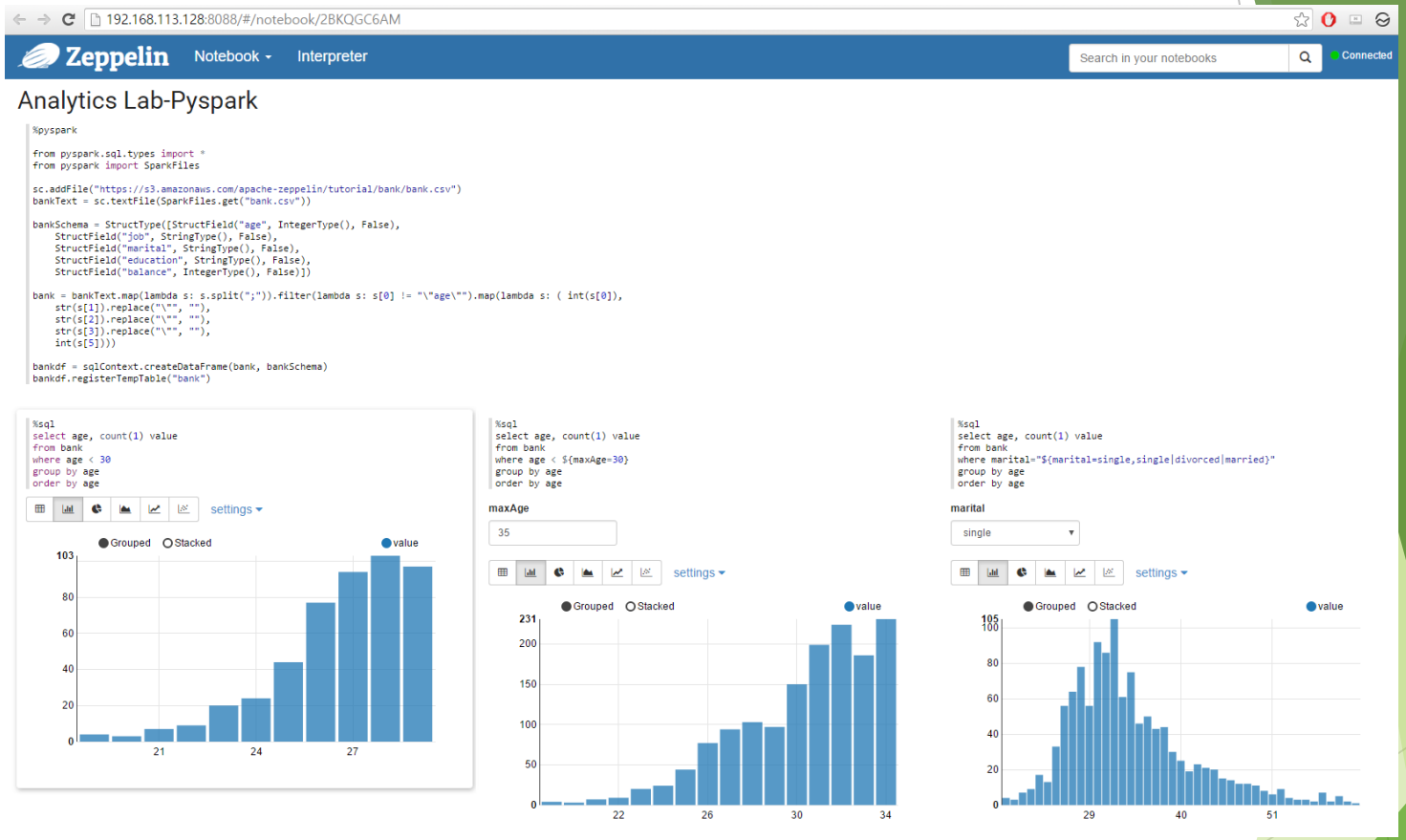
class pyspark.RDD

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

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

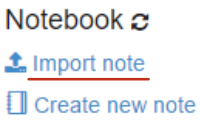
2. Background

- Pyspark: Spark in Python Language



Zeppelin tutorial converted to pyspark (https://github.com/SmartX-Labs/Mini/blob/master/Lab-7.%20Analytics/Analytics_Lab-Pyspark.json)

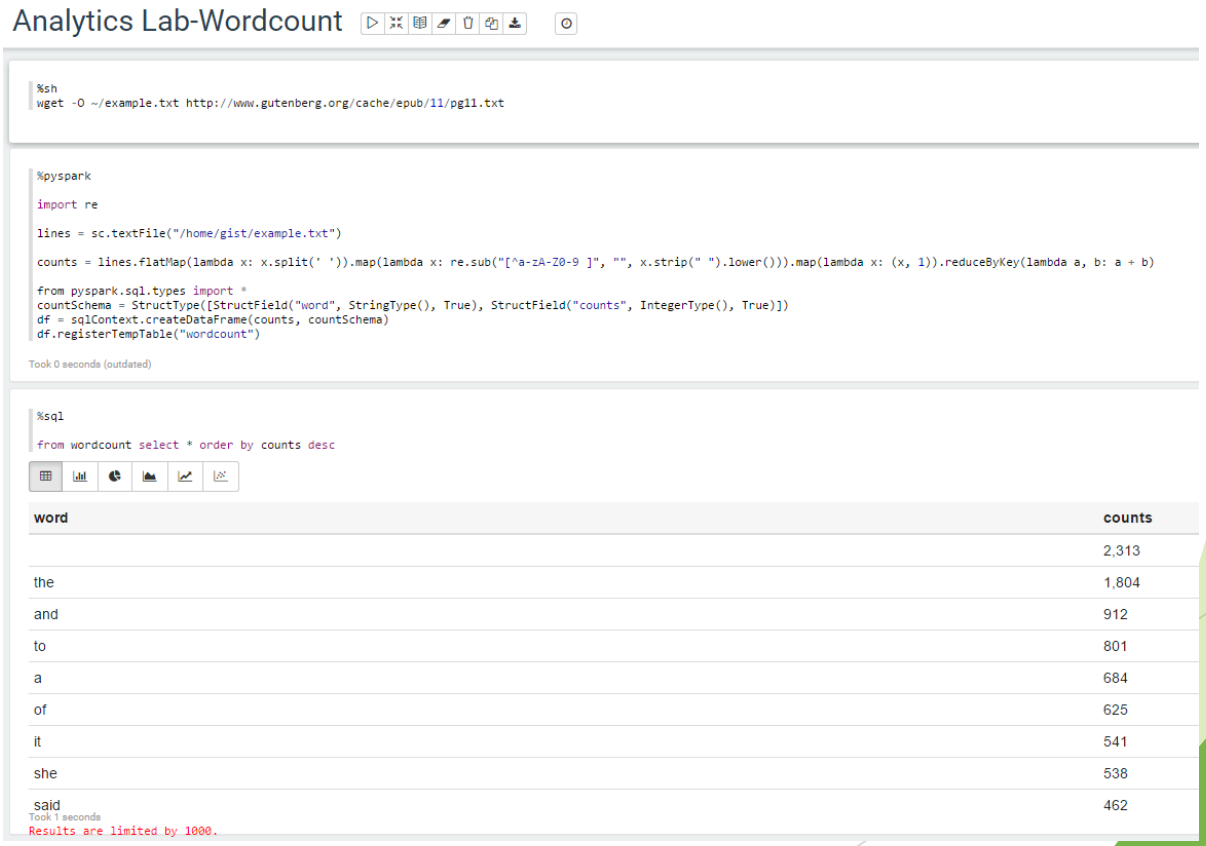
- In Import note and run.



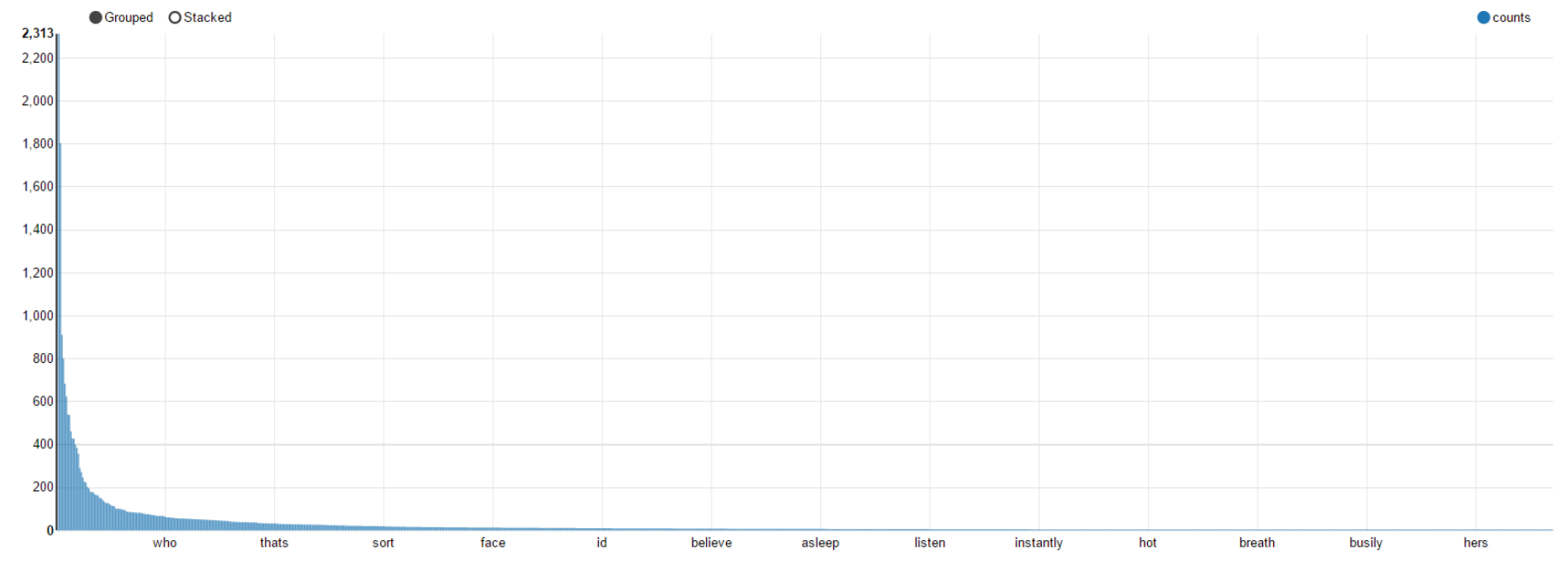
3. Apache Zeppelin
- Wordcount

Download and import notebook:
https://github.com/SmartX-Labs/Mini/blob/master/Lab-7.%20Analytics/Analytics_Lab-Wordcount.json

Change path of example.txt



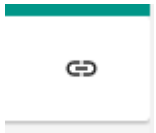
3. Apache Zeppelin
- Wordcount (Result-Visualization)



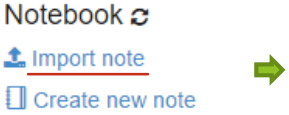
3. Apache Zeppelin
- Explore Zeppelin notebooks

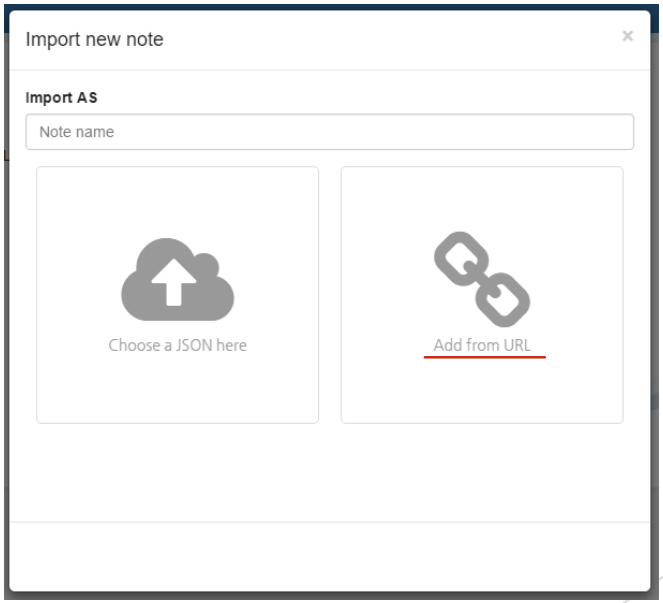
https://www.zeppelinhub.com/viewer

Look around the interesting notebooks.
And try to run on your machine by importing a notebook.



Copy url





Thank You for Your Attention Any Questions?