Dynamic Apache Spark cluster for economic studies

Symposium on Nuclear Electronics and Computing - NEC'2019

SIMPLE

Iuliia Gavrilenko (REU Plekhanova, speaker)
Tatyana Tikhomirova (REU Plekhanova)
Mayank Sharma (CERN)
Maarten Litmaath (CERN)

Prerequisites

diversity of methods for analyzing economic data that should be applied while complementing each other

huge data volumes, incompleteness and heterogeneity of the initial information

process of analyzing economic data is not trivial

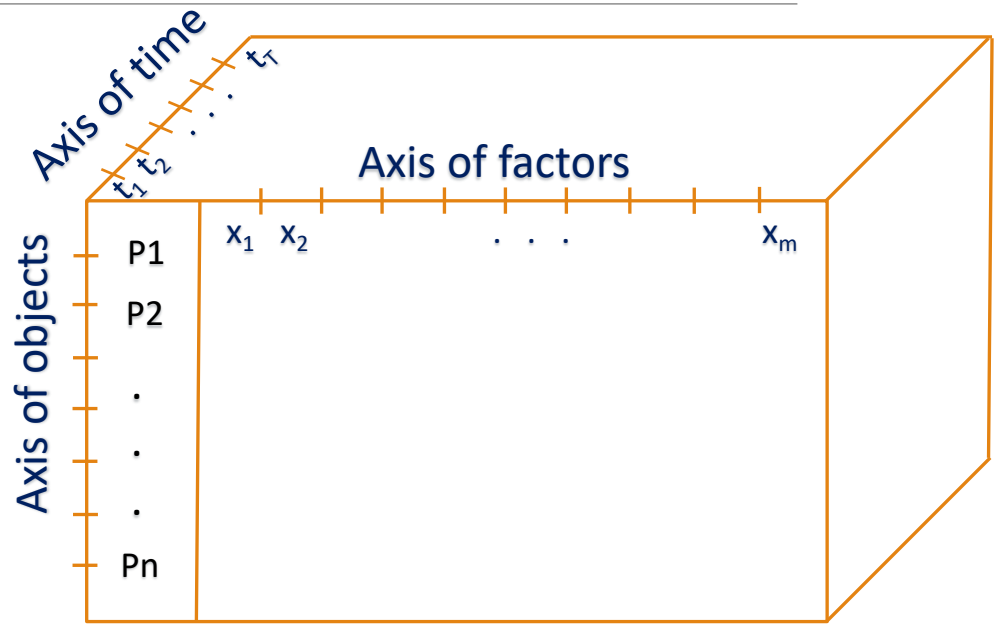
- SIMPLE Grid: http://cern.ch/go/8JLH

The level of employment in the regions of the Russian Federation

from

3 vectors of macroeconomic data OLAP cube:

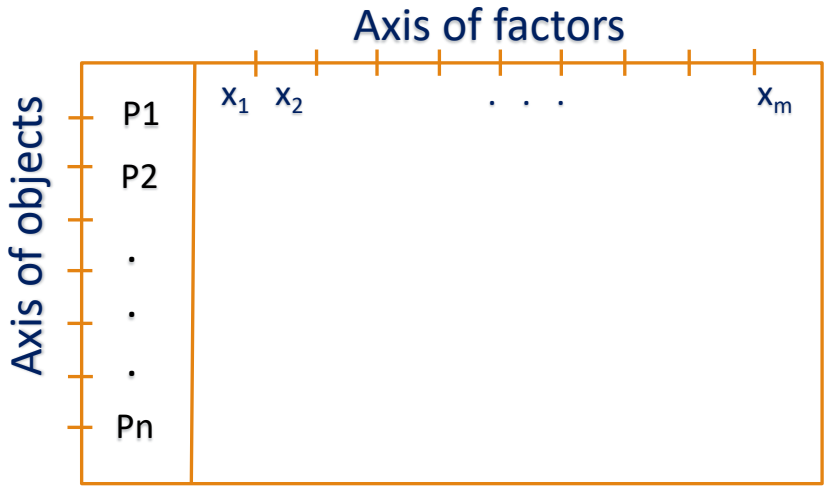
- axis of factors:
  - per capita income of the population (RUB);
  - volume of paid services per capita (RUB);
  - the cost of a fixed set of consumer goods and services (RUB);
  - the rate of migration increase in population (people);
  - the average size of the assigned pensions per capita (RUB);
  - turnover per capita (RUB);
  - actual final household consumption per capita (RUB);
  - industrial production index (%);
  - need for employees, declared by employers to employment service agencies (people);
  - fixed investment (RUB);
- axis of objects (Russian Federation regions);
- axis of time (years, months).



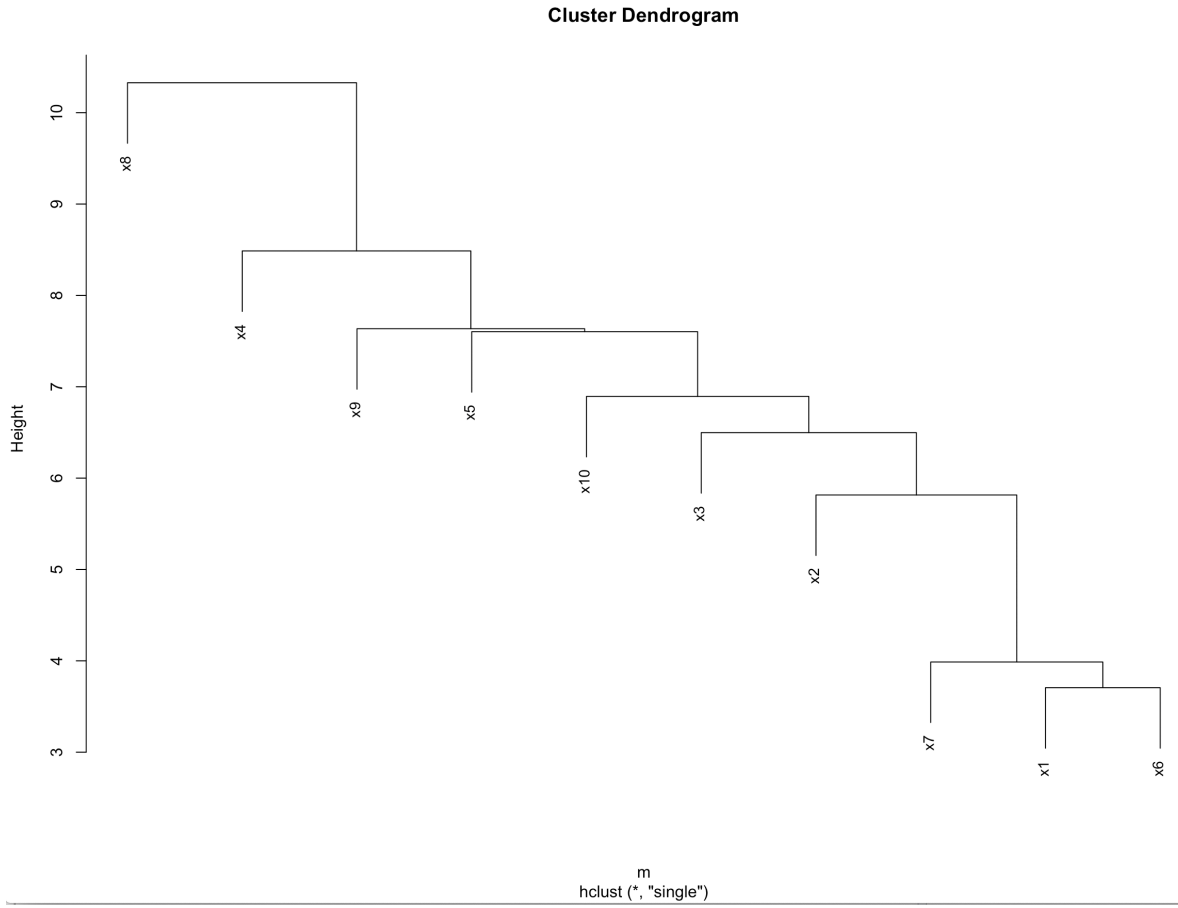
The level of employment in the regions of the Russian Federation

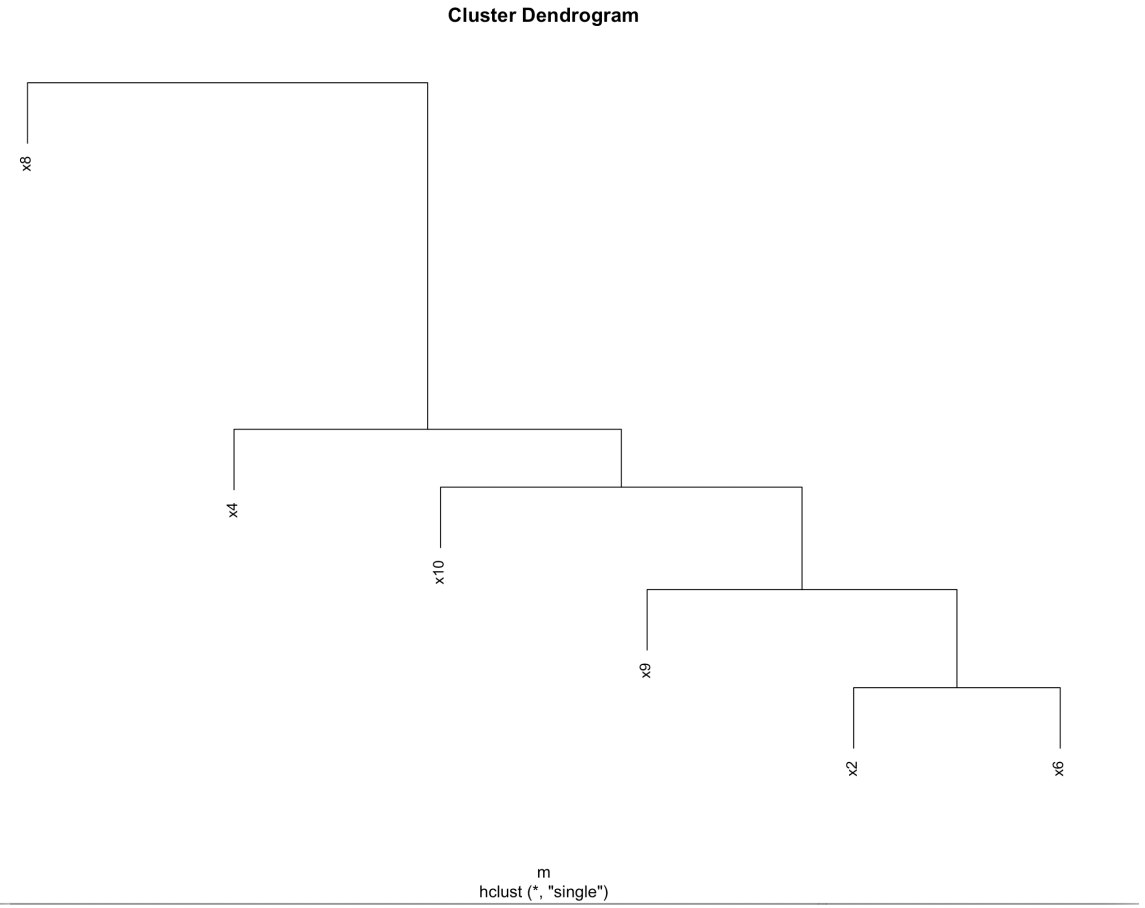
to macroeconomic data matrix:

- axis of factors:
  - per capita income of the population (RUB);
  - volume of paid services per capita (RUB);
  - the cost of a fixed set of consumer goods and services (RUB);
  - the rate of migration increase in population (people);
  - the average size of the assigned pensions per capita (RUB);
  - turnover per capita (RUB);
  - actual final household consumption per capita (RUB);
  - industrial production index (%);
  - need for employees, declared by employers to employment service agencies (people);
  - fixed investment (RUB);
- axis of objects (Russian Federation regions);

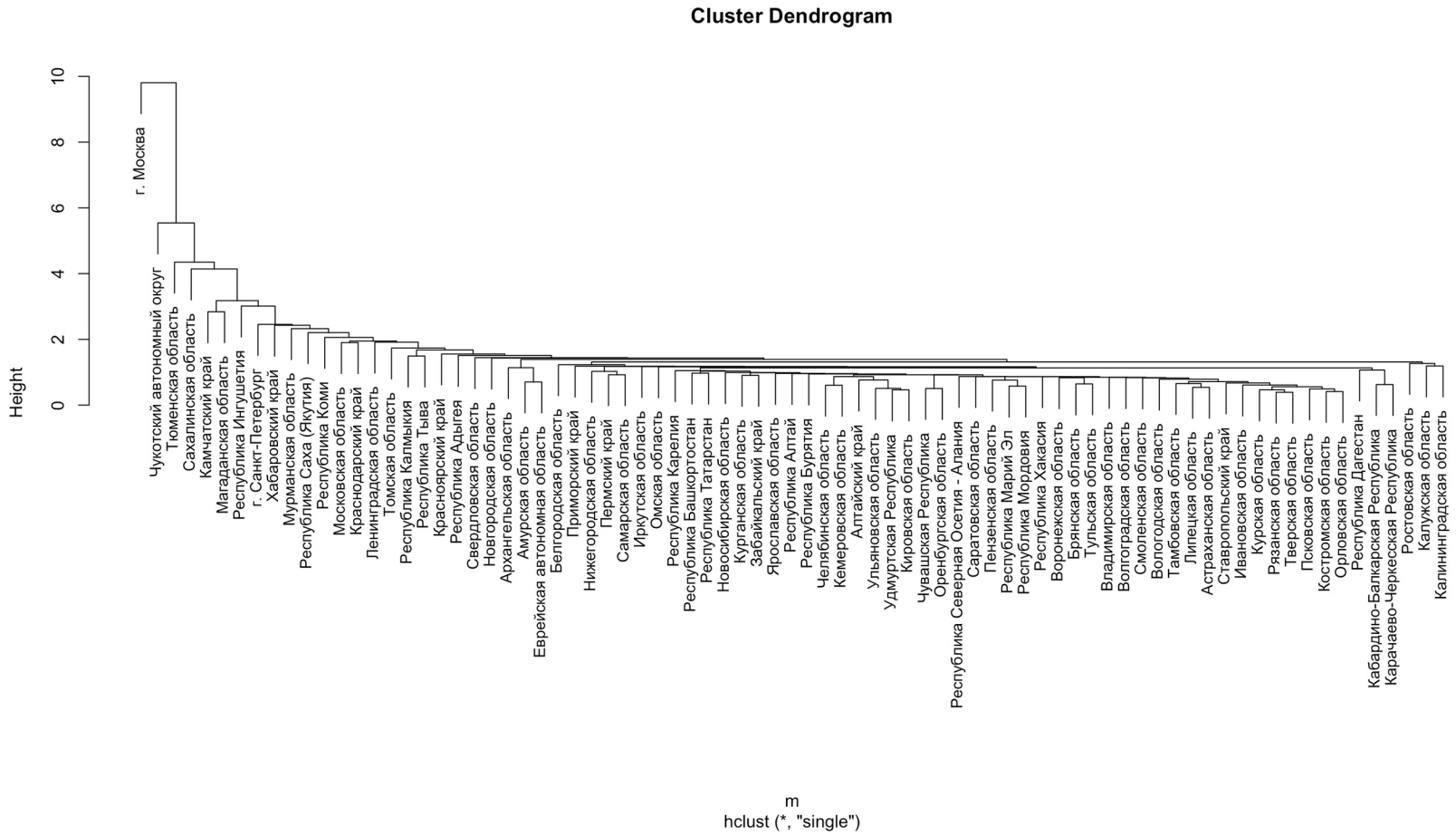


Clustering: single linkage by factors

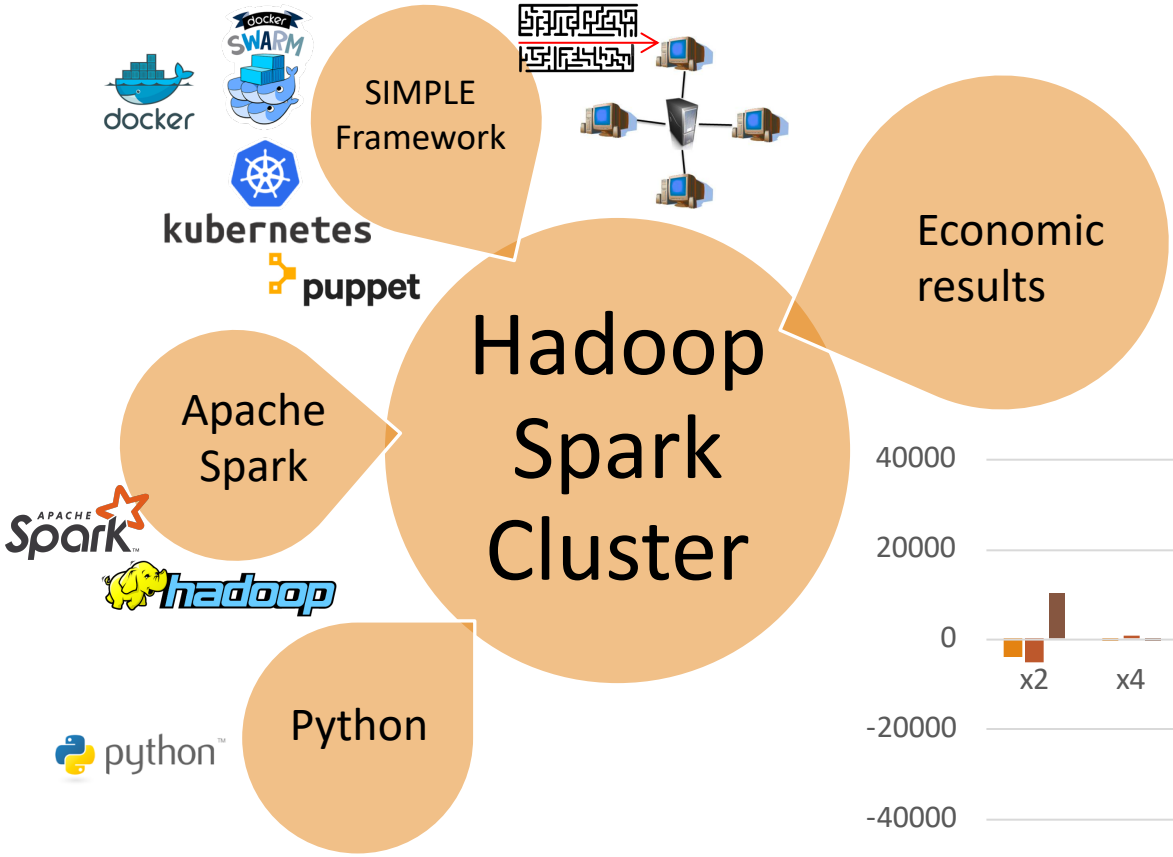


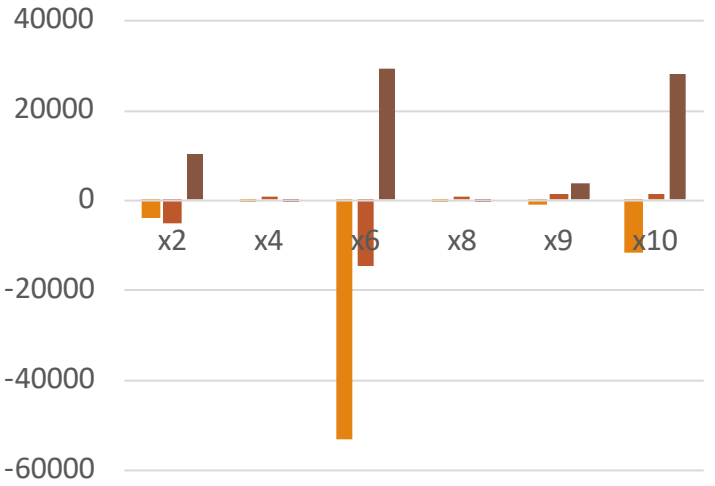


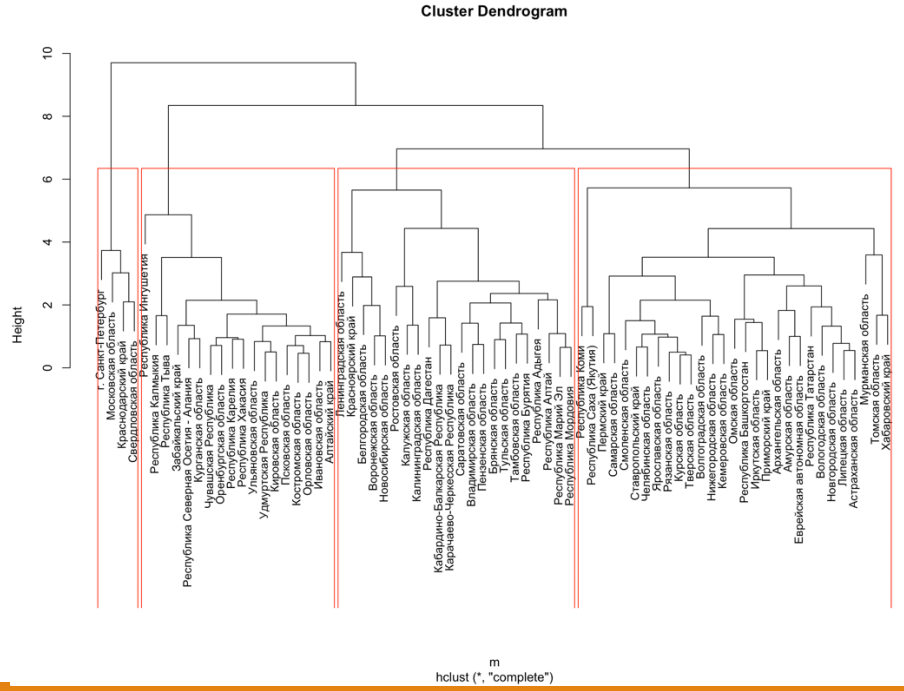
Clustering: single linkage by regions



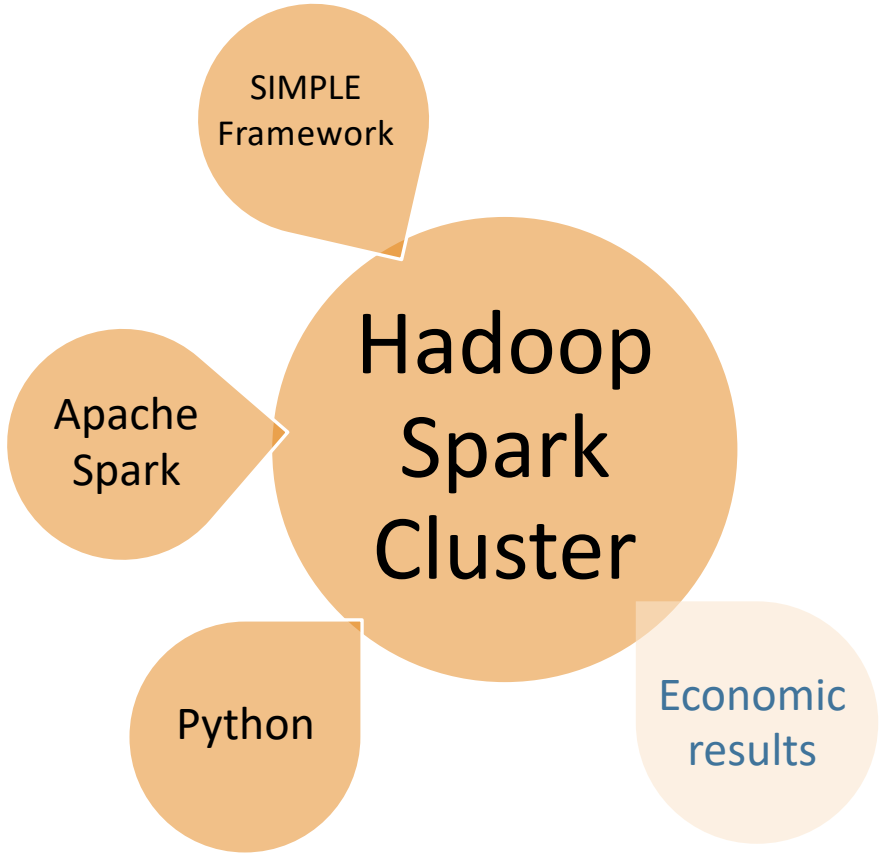
Dynamic Apache Spark cluster







SIMPLE: Project Structure



SIMPLE - **S**olution for **I**nstallation, **M**anagement and **P**rovisioning of **L**ightweight **E**lements

Support diversity in WLCG sites with **minimal oversight and operation efforts**

Keep **functionality the same**, but easier for site admins to setup and maintain

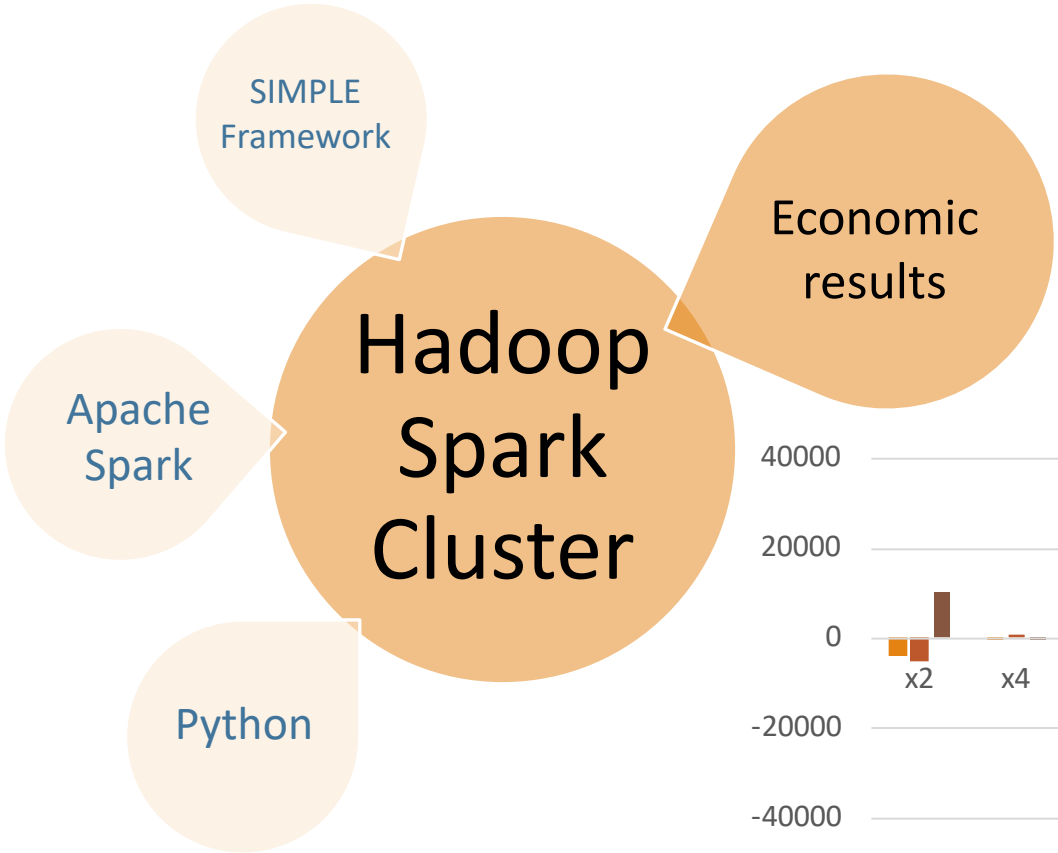
Implementation 2:

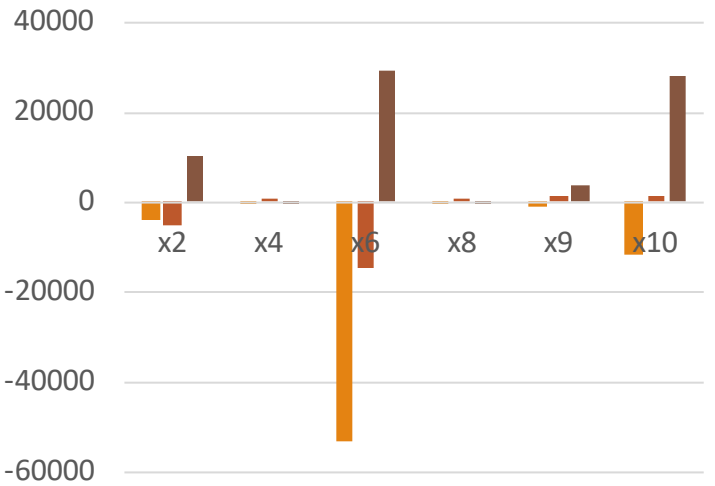
Site infrastructure: Docker + Swarm, Ansible (Implementation 1: Puppet; Upcoming: Docker + Kubernetes)

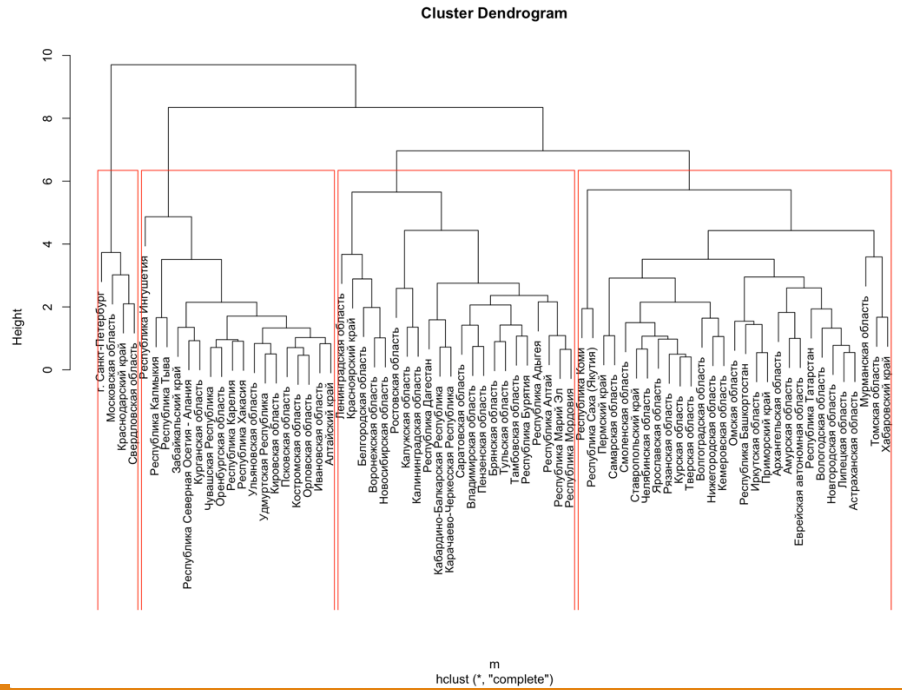
Hadoop Spark cluster for economics analysis:

- The SIMPLE Grid project is being developed by Mayank Sharma, software engineer with WLCG at CERN

Dynamic Apache Spark cluster: Economic results







Dynamic Apache Spark cluster

Небезопасно — spark-hadoop-master-0.cern.ch

Home | Nodes of the cluster | Namenode information

Logged in as: dr.who

Nodes of the cluster

- Cluster
  - About
  - Nodes
  - Node Labels
  - Applications
    - NEW
    - NEW_SAVING
    - SUBMITTED
    - ACCEPTED
    - RUNNING
    - FINISHED
    - FAILED
    - KILLED
  - Scheduler
- Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
6	0	1	5	3	4.50 GB	15 GB	0 B	3	40	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
5	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[MEMORY]	<memory:256, vCores:1>	<memory:3072, vCores:4>	0

Show 20 entries Search:

Node Labels	Rack	Node State	Node Address	Node HTTP Address	Last health-update	Health-report	Containers	Mem Used	Mem Avail	VCores Used	VCores Avail	Version
	/default-rack	RUNNING	spark-hadoop-add-1-worker.cern.ch:34042	spark-hadoop-add-1-worker.cern.ch:8042	Fri Apr 19 13:25:31 +0200 2019		1	1.50 GB	1.50 GB	1	7	2.8.5
	/default-rack	RUNNING	spark-hadoop-submit.cern.ch:45121	spark-hadoop-submit.cern.ch:8042	Fri Apr 19 13:25:31 +0200 2019		0	0 B	3 GB	0	8	2.8.5
	/default-rack	RUNNING	spark-hadoop-add-3-worker.cern.ch:35466	spark-hadoop-add-3-worker.cern.ch:8042	Fri Apr 19 13:25:31 +0200 2019		0	0 B	3 GB	0	8	2.8.5
	/default-rack	RUNNING	spark-hadoop-worker-0.cern.ch:44695	spark-hadoop-worker-0.cern.ch:8042	Fri Apr 19 13:25:31 +0200 2019		0	0 B	3 GB	0	8	2.8.5
	/default-rack	RUNNING	spark-hadoop-add-2-worker.cern.ch:40365	spark-hadoop-add-2-worker.cern.ch:8042	Fri Apr 19 13:25:31 +0200 2019		2	3 GB	0 B	2	6	2.8.5

Showing 1 to 5 of 5 entries First Previous 1 Next Last

swan003.cern.ch

Nodes of the cluster | Namenode information | CERNBox | LR

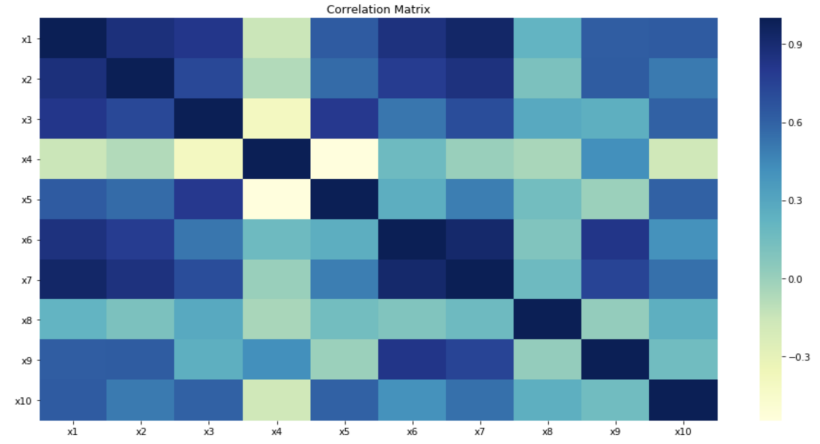
FILE EDIT VIEW INSERT CELL KERNEL WIDGETS HELP Not Trusted Python 3

Code

Plots

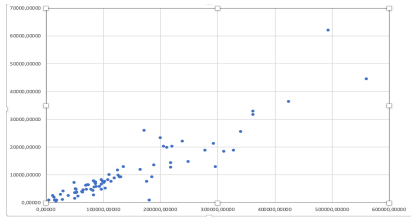
In [40]:
```
from pyspark.ml.stat import Correlation

matrix = Correlation.corr(train_df.select('features'), 'features')
matrix_np = matrix.collect()[0]["pearson({})".format('features')].values
```

In [43]:
```
import seaborn as sns
import matplotlib.pyplot as plt

matrix_np = matrix_np.reshape(len(['x1','x2','x3','x4','x5','x6','x7','x8','x9','x10']),-1)

fig, ax = plt.subplots(figsize=(13,7))
ax = sns.heatmap(matrix_np, cmap="YlGnBu")
ax.xaxis.set_ticklabels(['x1','x2','x3','x4','x5','x6','x7','x8','x9','x10'], rotation=0)
ax.yaxis.set_ticklabels(['x1','x2','x3','x4','x5','x6','x7','x8','x9','x10'], rotation=0)
ax.set_title("Correlation Matrix")
plt.tight_layout()
plt.show()
```

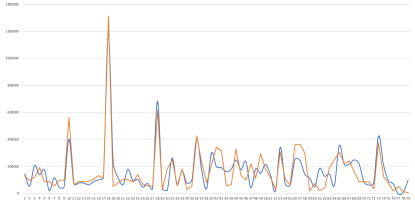


In [44]:
```
from pyspark.sql.functions import avg
```

Conclusions



The degree of tightness of the linear relationship between the main macroeconomic factors was measured



Constructed several models of the regional labor market, taking into account the main macroeconomic factors

Was obtained classification of the subjects of the Russian Federation according to the level of employment

A dynamic Apache Spark cluster deployed by the means of the SIMPLE framework developed at CERN was used for all analysis

The Community

GitHub Repositories

- https://github.com/JuliaGavrilenko/site_level_config_file
- https://github.com/JuliaGavrilenko/simple_spark_cluster_master

SIMPLE Project Website

Link: http://wlcg-lightweight-sites.github.io

Open Source Community

Name: WLCG Lightweight Sites

Link: http://cern.ch/go/Hz7S

Simple Grid Specification

https://docs.google.com/document/d/1yp_96UXcwNO49cktnHtT61iNmTO0RgrSQukuNYqACpM/edit?usp=sharing

Thank you for your attention!