On the IT infrastructure of Russian Academy of Sciences

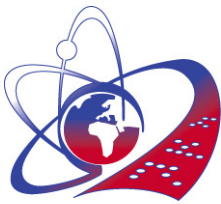
Heading to the Data Intensive Science

Lev Shchur

Science Centre in Chernogolovka

Outline

- e–infrastructure of SCC
- Ways of development
- Data Intensive Science for …
- Big Data – myth and reality
- e-realization of DIS
- DIS in research institutes



e–infrastructure *of SCC RAS*

1) Network ChANT (**Ch**ernogolovka **A**cademic **Net**work);

2) e-library;

3) FTP server with Open software;

4) Cluster WALL;

5) Cluster Manticore;

6) Cloud «Тучка»;

7) Video conferencing system VideoGrid;

8) Monitoring and management.

Tendency of e–infrastructure development

- Data Intensive Science
- Big Data
- e–infrastructure – testbed for IT development, new technologies implementation, …
- Atlas – Grid, …
- Federal Tax Agency
- Distributed Hardware & Centralized Software with the goal - Big Data on the working table

(Very) Big Data = V^4

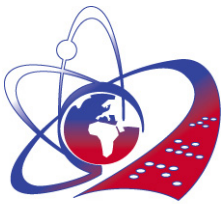
- *Volume – very big data volume;*
- *Velocity – big speed of data processing;*
- *Variety – big diversity of data;*
- *Veracity – data veracity.*

Big Data – myth and reality - 1

- e–resources *analysis – 29 NRU RF*
- e–resources *analysis – Institutes in Physics, RAS*
- e–resources *analysis – leading Universities, USA and EU*
- e–resources *analysis – National Research Labs, USA and EU*
- e–resources *analysis – Intl projects*

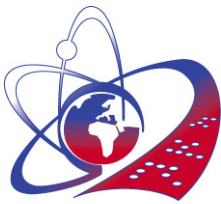
Big Data – myth and reality – 2 (RF)

- NRU – *only ½ have some data*
- *Interesting data - MSU, HSE, NNU, NGU, Miners, ITMO. Mainly sociology.*
- *½ NRU – e–libraries*
- *RAS Institutes - e–libraries, astrophysics, space research, meteorology*



Big Data – myth and reality – 3 (West)

- *US Universities – CalTech: genome, brain, geodynamics, space research*
- *Mainly – local resources, no open access data*
- *Mainly – public relation, small amount of data*
- *Intl projects – only for the members of collaborations*

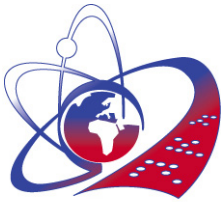


Scientific Journals – myth and reality – 4

- 22 RF journals – full texts (**Письма в ЖЭТФ, УФН**, Труды МИАН, ПМТФ, Проблемы передачи информации, Математические заметки, ЖЭТФ)
- 15 Intl Journals – all volumes (Philosophical Transactions, Nature, Physical Review, Chemical Reviews, Science, …) – not open access
- Письма в ЖЭТФ, УФН **– all and open**

What one can gain from Big Data?

- *Извлечение смысла из больших данных (Data Intensive Science)*
  - *LHC-Grid project (V 3)*
  - *IBM Watson project (V 4)*



e–infrastructure and Big Data

Problems and features? - 1

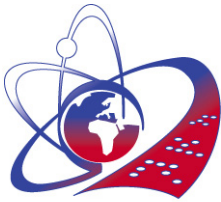
- Data store and transmit – Big Data Volume (V#1)
- Data processing – Big Speed (V#2)
- Visualization – Big Resolution (V#2)

HardWare requirements!

Problems and features? - 2

- Data Processing – data diversity, unformated data, diversity of representations, text on many human languages (V#3)
- Data Veracity – data reliability (V#4)

SoftWare requirements!

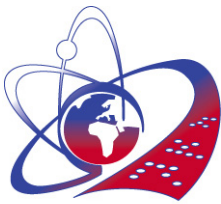


e–infrastructure – Technical View

- 1) Communication channels;
- 2) Telecommunication centers – connection with the networks inside RF;
- 3) Telecommunication centers – connection with the networks inside RF;
- 4) HW for Data Storage;
- 5) HW for Data Processing;
- 6) HW for Data Transmission;
- 7) HW for User Access.

e–infrastructure – Functions

- 1) Data Exchange within Research Collaborations;
- 2) Distributed Scientific Conferences and Workshops;
- 3) Research in Computer Sciences;
- 4)) Storing, Processing, Transmitting and Accessing Big Data for DIS



e–infrastructure for sciences – Features

- 1) Network Policy for Research Collaborations;
- 2) Network Policy for Internetwork Exchange;
- 3) Testbed for Scientific Experiments and for Emerging Technologies

Предложения по комплексу работ

- 1) использовании имеющейся ИКТ инфраструктуры учреждений ФАНО для проведения фундаментальных исследований в области Больших данных;
- 2) разработки программы научных исследований в области Больших данных для ее выполнения силами научных коллективов учреждений ФАНО;
- 3) разработки программы мероприятий по внедрению сервисов работы с Большими данными.
- 4) наипервейшего решения требует проблема бюджетного финансирования магистральных каналов.

Пути решения (конкурсы, субсидии, …)

- Распределенная аппаратная часть е-инфраструктуры
- Централизованные программные интегрирующие системы
- Специализированные системы обработки Больших Данных
- Дружественный интерфейс пользователя

Пути решения

цитата из закона о связи:

“Технологические сети связи предназначены для обеспечения производственной деятельности организаций, управления технологическими процессами в производстве.”