ICHEC
Irish Centre for High-End Computing

The UNECE Big Data Sandbox: What Means to What Ends?

Bruno Voisin
Niall Wilson

UNECE Workshop on the Modernisation of Statistical Production
Geneva 15th-16th April 2015

The UNECE Big Data Sandbox: Mission

- Test the feasibility of remote access and processing.
- Test existing statistical standards/models/methods in a Big Data context.
- Determine which BD tools are most useful to statistical organisations.
- Gain experience on BD handling and processing.
- Build an international collaboration community to share ideas and experience on BD.

The Sandbox, then and now

20 nodes, each:

- 2xIntel Xeon X5560 quad-core processors
- 48GB RAM
- 1x1TB disk
- DDR Infiniband (16Gbit)

May 2015

4 nodes, each:

- 2xIntel Xeon E5-2650 v3 10 core processors
- 128GB RAM
- 4x4TB disk
- FDR Infiniband (56Gbit)

?

Hortonworks

hadoop

STORM

Spark

Lightning-Fast Cluster Computing

R

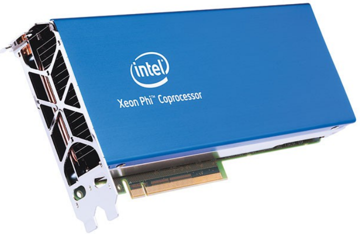
Hardware Technologies

x86 CPUs

GPUs

Large Shared Memory

Manycore



FPGAs

Burst Buffer Cache

INFINITE MEMORY ENGINE™

Software Technologies

Hadoop ecosystem:

- Hbase
- Storm
- Spark

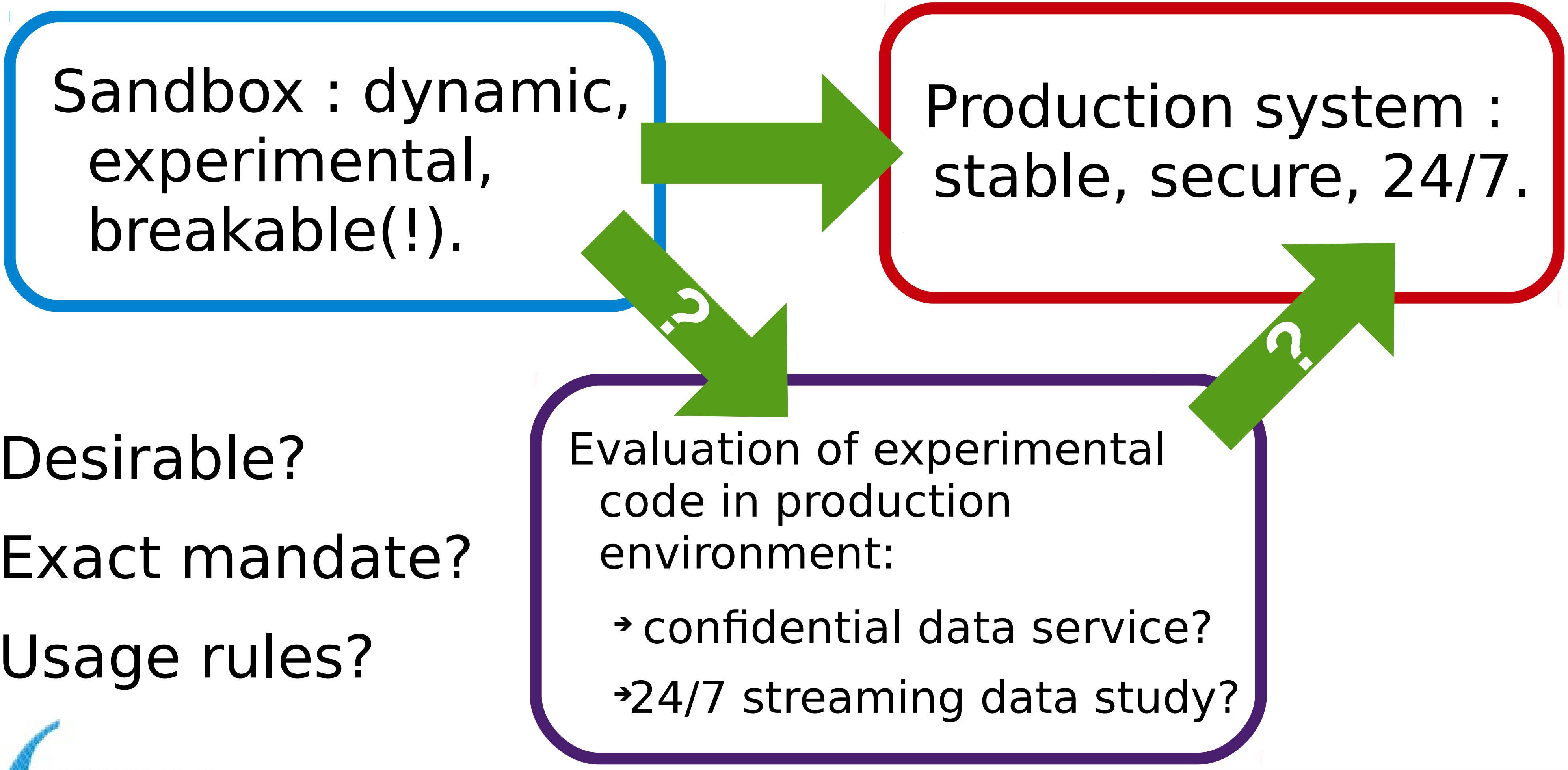
In-memory DBs:

- SAP HANA
- MonetDB
- kdb+
- ...

HPC approach: MPI.

R as an interface to everything on this slide!

Broadening the Mandate: (pre-)Production Environment



Sandbox : dynamic, experimental, breakable(!).

Production system : stable, secure, 24/7.

Evaluation of experimental code in production environment:

- confidential data service?
- 24/7 streaming data study?

Desirable?

Exact mandate?

Usage rules?

Broadening the Mandate: Synthetic Data

Real/Confidential data:

- Data not ready
- Limited partners
- legal requirements

Synthetic data:

- Ready and at-scale
- Available to anyone
- No lawyers involved!

Easy third-party collaborations (ex: ICHEC with CSO and ESRI).

More systematic use of synthetic data?

Shared data generation facility?

Conclusions / Perspectives

Current small Hadoop-based Sandbox is a first step.

Fitting use cases should take advantage of it, other cases should guide the thinking forward.

What will the statistical community want to do in the near future? Production environment and third party collaboration through synthetic data may be part of the answer, but not all of it.