Contribution ID: **21** Type: **Full presentation: short (15 mins.)**

Data Challenges at the Square Kilometre Array (SKA)

Tuesday, 3 November 2020 14:00 (15 minutes)

The upcoming observatory Square Kilometre Array (SKA) transports *data products* to SKA Regional Centers (SRC). Realizing a global SRC network is of utmost importance for radio astronomy and opens, beyond that, unique opportunities for developing generic infrastructure components that are of interest for other communities as well.

The resolution power of sensors is increasing steadily resulting in larger and larger data volumes. For reasons of sustainability, everybody is getting sooner or later to the point that only a tiny fraction of the generated data can be stored in the long term.

Experiments archive their data and analyze them over an over again. This traditional method resulted in unexpected discoveries. For example, "Fast Radio Bursts (FRB)" are high-luminous signals from rare cosmic events that were detected in 2007 by analyzing data taken a few years earlier in 2001. The raw data volumes at SKA are so large that only a small fraction can be stored in archives. The necessary strong data reduction has to be provided nearly in real-time. This very time constraint results in fundamental challenges.

1. Data Irreversibility
   Experiments archive all data, in general, in order not to loose any information. The real-time constraint, however, limits the effectivity of this approach simply because there is not enough time to process workflows in full detail. Missing information cannot be recovered later on whereby an "arrow of time" is introduced, an essential characteristic of irreversibility. To reduce irreversibility effects, feedbacks should be integrated into the global workflow. Firstly, the outcome of a "fast analysis" of online data could be used to optimize the control of sensors.
2. Dynamic Archives
   Archived datasets have to be characterized by quality measures that are adjusted regularly based on simulations. The outcome of this "slow analysis" could be used for steering the sensors via a further feedback. In other words, archives will no longer be static but dynamic entities. Accordingly, metadata schemes should be extendable dynamically to keep up with increasing knowledge.
3. Data Monster
   The great number of antennas at SKA provide images from the cosmos of unprecedented resolution. Single images may be as large as one Petabyte. Analyzing objects of such size requires a shift in paradigm: from currently processor-centric to memory-based computing architectures. It should, however, be noted that further efforts are needed. Speedup in parallel computing relies on the Divide&Conquer principle which, in turn, is based on the assumption that the problem class of a split dataset is equivalent to the original dataset. Medical image processing indicates that each Divide&Conquer-step may need a careful justification.

In the presentation, the impact of the three challenges on future data infrastructures is elucidated, in the general as well as on the global SRC network of SKA, based on discussions within the German SKA community, which is organized by the "Association of data-intensive Radio Astronomy (VdR)". The connection to related work is clarified, e.g. to the concept of "data lakes" in high-energy physics, and to the outcome of the Big Data and Exascale Computing (BDEC) project.

Primary authors: Prof. HESSLING, Hermann (Verein für datenintensive Radioastronomie (VdR), and University of Applied Sciences (HTW) Berlin); Prof. KRAMER, Michael (Verein für datenintensive Radioastronomie (VdR), and Max Planck Institute for Radio Astronomy (MPIfR) Bonn); Prof. WAGNER, Stefan (Verein für datenintensive Radioastronomie (VdR), and Zentrum für Astronomie Heidelberg (ZAH))

Presenter: Prof. HESSLING, Hermann (Verein für datenintensive Radioastronomie (VdR), and University of Applied Sciences (HTW) Berlin)

Session Classification: Cloud computing - Part 1