Data Processing in Radio Astronomy - Preparing for the Square Kilometre Array

White Paper for BDEC2 meeting in Poznan – May 2019

Prepared by:
M.P. van Haarlem (ASTRON), J. van Leeuwen (ASTRON/Amsterdam), J.B.R. Oonk (SURFsara/Leiden/ASTRON),
T. Shimwell (ASTRON/Leiden), L.V.E. Koopmans (Groningen)

This document summarizes three use cases of computationally intensive LOFAR[1] data processing and analysis workflows. These demonstrate the kind of processing required for the Square Kilometre Array (SKA) radio telescope, which is expected to be fully operational from 2026/2028. The workflows presented all require further development and new services in order to deal with the increased data volumes expected in the operational phase of the SKA. These developments strongly challenge the convergence between HPC, HTC and HDA data and computing capabilities, and their integration in the context of a software platform of distributed services across a continuum of edge and centralized (Cloud, HPC) infrastructures to support such complex, wide-area workflows and their data logistics.
In the operational phase, SKA will involve multi-stage, and geographically distributed science-driven data processing and analysis workflows. The first stage in the two telescope host countries (Australia and South Africa) will deal with the processing and reduction of raw multi-source data streams across edge (close to the instruments) and centralized (HPC) infrastructures to deliver first stage (Observatory) multi-type Data Products – i.e., event and visibility data. These primary data products will be distributed to a network of ~10 geographically distributed SKA Regional Science Centres[2] that archive the data and where the second stage of the science application high-end data processing/analysis takes place. The total data volume transported to and archived in these regional centres will grow to a rate of ~1 Exabyte per year from ~2027/8 and raise challenging and potentially disruptive issues in terms of data logistics, i.e. the management of the time sensitive positioning and encoding/layout of data, relative to its intended users and the computational and data resources and services that they can apply. In many countries the SKA Regional Science Centres will make use of common data and computing infrastructure shared with other data/compute intensive disciplines.

The H2020 ESCAPE[3] is a first step towards the collaboration between partners from astronomy and particle physics in the context of the European Open Science Cloud (EOSC). ESCAPE aims at delivering solutions to ensure integration of data, tools, services and scientific software; to foster common approaches to implement open and FAIR data services and stewardship; to establish within EOSC interoperability services enabling an integrated multi-probe facility for fundamental science. Integration of a continuum of existing European edge and centralized (HPC, Cloud) infrastructures into a software platform of distributed services will foster the convergence between HPC, HAD and HTC (including data-driven Machine Learning and AI), and shall provide an enormous boost to the discovery potential through the interoperability of planned large science-driven infrastructures. As such this can contribute to a science-driven demonstrator for BDEC2 shared not only between astronomy (SKA) and particle physics (CERN/LHC) communities but also other communities such as Space Observation.

Disruptive issues in the coming decade for LOFAR/SKA:

1) Any large scale (KSP) type projects, with a semi-continuous data flow, must move to a workflow solution on compute near the data. This to some degree means professionalizing the current codes, embracing CI/CD developments to orchestrate update and community input, and it means giving up (some of) the freedom of doing data reduction on your own laptop/mini-cluster.

---

[1] https://www.astron.nl/telescopes/lofar
[2] Design activities for these regional centres are ongoing in most of the SKA member countries. In Europe these are being coordinated in the H2020 AENEAS project (www.aeneas2020.eu).
[3] ESCAPE brings together the following ESFRI facilities (CTA, ELT, EST, FAIR, HL-LHC, KM3NeT, SKA) as well as other pan-European research infrastructures (CERN, ESO, JIV-ERIC, EGO-Virgo) in the astronomy and particle physics research domains.

2) Very high memory imaging. We will breach the 20x20k image size in the next decade and move towards 100kx100k images. With current DD(F) pipelines this will require about 2 TB of memory on a single node to run efficiently. New types of parallelisation which can spread the DD(F) calibration/imaging load across multiple nodes may lower this requirement.

3) High levels of automated quality control to lower the pipeline failure rate - however this may mainly affect the initial calibration stages. ML-assisted auto-tuning may play a role here, but it still needs to be developed. To what extend these methods may also play a role in the later DD calibration/imaging is not yet clear (i.e. can optimised cal/img strategies be found based on (i) data and (ii) astronomers reqs (e.g. diffuse emission and combining it with high resolution emission - how to (simultaneously) image scales that span more more than 2 orders in magnitude in baseline length?)

4) On-demand resources in terms of fat nodes for radio astronomy are very expensive. Tuning the workload and the defining interaction models between astronomers and the underlying hardware needs to be investigated (cloud vs. batch, or both?)

5) Grid and X509 solutions must be updated to cloud (probably managed cloud service layers and not self-service) and AAI solutions in the coming years - this requires coordination across the edge to continuum of resource providers, observatories and user(institutes).
Data storage models will also change to data lake-type models. This will require a new layer of organisation or alternatively abstraction. Similarly workflows will need to run in a seamless manner across infrastructures, which implies that they either need to be independent of underlying or be made aware via intelligent middleware layers.

6) The radio astronomy case, in particular SKA, will break open a new realm in compute, that is fundamentally different from HEP in terms of data organisation and processing requirements. Whereas HEP workflows consist of many small event files that can be treated independently, the opposite is true for the SKA in that a through complex set of calibration tasks all data needs to converge at several instances in this workflow, i.e. the data is not independent (although some parts of it can be treated via diverging flows before ultimately converging again).

Case 1: Pulsars

Pulsars are fast spinning, highly magnetized neutron stars that provide physical conditions far beyond those attainable in Earth-bound laboratories. Studying their extremes is thus key to understanding fundamental gravitational and particle physics. The SKA will capitalise on digital signal processing and computing power to provide an unparalleled field-of-view. The SKA will survey the sky for their periodic, single-pulses (Fast Radio Bursts) and slow-transient emission.

Finding radio pulsars in radio-telescope data is exceedingly data and computationally intensive. Data streams from the SKA stations and dishes are first pre-processed on dedicated edge FPGA hardware. The resulting streams are combined in beamformers on site and in centralised facilities in the host countries. Data streams are split into independent pointings that need to be further processed and analysed individually outside the host countries in a network of regional centres.

We will be performing massive surveys with SKA; for both their periodic, single-pulse (Fast Radio Burst) and slow-transient emission. These searches are a massive data and computational challenge across a continuum of edge and centralised (HPC, Cloud) infrastructures raising disruptive issues in term of distributed data and computing services, as well as in term of data logistics across this continuum.

<u>Pulsar Processing Details</u>

1) Data preparation – creation of filterbank files

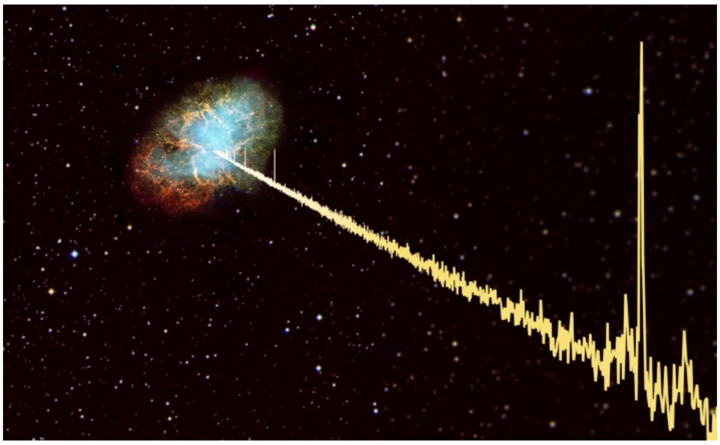
This is a highly disk I/O–intensive process, where many >10-GB size files are read from a scratch disk pool or from the incoming network, re-sampled, and written back to shared storage. This requires a central data handling location on the input side of the HPC/Cloud centralised facility.

2) Data preparation – Radio frequency interference (RFI) excision
In interference excision, samples in a 2D matrix that strongly deviate from the mean are identified and removed. This is a more computationally expensive process, where all data is read and analysed in order to create an RFI mask. The (current) best performance is achieved by embarrassingly parallel farming over nodes.

3) Search for periodic signals and single pulses
The main algorithms for this stage are:

a. Dedispersion (progressively shifting all rows in a 10 GB 2D matrix, and collapsing to 1D)
b. Discrete Fast Fourier Transforms (FFTs)
c. Matched filtering – both in the time and frequency domain
d. Folding (chopping the data into similarly sized chunks and adding, algorithmic sped up using FFAs)
e. Machine learning algorithm for candidate scoring and classification

For pulsar and FRB surveys with the LOFAR (Stappers et al. 2011; Sanidas et al. 2019) and Apertif (Maan & van Leeuwen 2017) telescopes, pathfinder instruments for the SKA, the processing in steps 1-3 amounts to 1 and 5 PFLOPS respectively. In each, about 1 PB of data is recorded and processed per day. These are powered by the Dutch national supercomputer Cartesius, and by a dedicated GPU/CPU cluster. For SKA, the planned compute required is of order 100 PFLOPS.



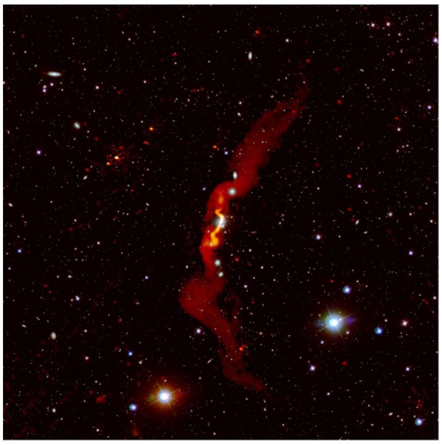


Figure 1 (left): A 1-second snapshot of the radio peaks that the Crab pulsar emits. In the background the outer layers that were ejected in the supernova are visible in the optical (ASTRON Westerbork/ESO VLT composite image).

Figure 2 (right): The radio galaxy 3C31, observed with LOFAR by Heesen et al (2018), is shown in red on top of an optical image. LOFAR reveals the radio galaxy to be more than 3 million light years in size. Credit: Volker Heesen and the LOFAR surveys team.

References
Coenen et al. 2014, A&A, 570, 60
Maan & van Leeuwen, 2017, Proc. URSI GASS 2017
Stappers et al. 2011, A&A, 530, A80

Case 2: LOFAR Surveys Imaging Workflow

The Low Frequency Array (LOFAR) is a technology pathfinder for SKA-Low. The pipelines, developed by the Surveys KSP (SKSP), produce thermal noise limited images at 6 arcsec resolution (see Shimwell et al. 2019) and are able to process data in a semi-continuous flow (on average 1 new dataset per day). The challenges in imaging and calibrating LOFAR data are similar to those that will be faced by SKA-low and as such the SKSP imaging pipelines provide a very good demonstrator for processing future SKA-Low continuum datasets. Furthermore, these pipelines have already proved capable of processing deep LOFAR observations with over 100hrs of LOFAR data.

The LOFAR surveys project will observe approximately 3,100 fields (1,300 already complete). Each field must be processed as described below.

1) Archiving the observations: LOFAR surveys data are recorded with a resolution of 64ch per 195 kHz subband and 1s integration times. This results in approximately 64TB of data per pointing and two pointings are observed simultaneously in an 8hr period and hence collecting ~128TB of data in 8 hrs. These data are flagged for interference and averaged to a resolution of 16ch per 195 kHz subband and 1s integration time and then ingested into the LOFAR archive. The resulting archived data has a size of 16 TB per pointing. This initial processing is completed on the CEP4 compute nodes and to date 450, 50, 800 datasets have been stored in a federated data archive that is hosted by SURFsara (Amsterdam), PSNC (Poznan) and FZ-Jülich respectively. In total the archived survey data will occupy approximately 30PB.

2) Correction of Direction-Independent (DI) Effects.: The archived LOFAR data must be calibrated to remove time independent (or slowly varying with time) instrumental effects such as the amplitude corrections required to get the flux scale correct, the offset between XX and YY phase and the clock offsets for each station. The procedure followed for this is described in van Weeren, et al. (2016), Williams, et al. (2016), Shimwell, et al. (2017) and de Gasperin, et al. (2019). The output data from this pipeline for each pointing are kept at a resolution of 2ch per 195 kHz subband and 8s integration time and occupy approximately 250GB and are stored on the SURFsara facilities. This processing is completed on the compute nodes at either SURFsara or FZ-Juelich (see Mechev et al. 2017) where the majority of the data are stored (data from Poznan are copied to SARA) and requires approximately 1000 CPU core hours per pointing. This step cannot be performed elsewhere due to limitations in transporting the large amounts of data.

3) Correction of Direction Dependent (DD) Effects: Low frequency radio data are severely corrupted by time and position dependent ionospheric errors which must be corrected in order to produce science quality images. Correcting these corruptions is still very much an area of active research. The surveys project uses a pipeline that makes use of *kMS* (Tasse 2014b) to calibrate for ionospheric effects and errors in the beam model and *DDFacet* (Tasse et al. 2017) to apply the derived solutions during the imaging. This pipeline takes about 5000 CPU core hours (5-7 on a single HTC node) days to image one survey pointing when operating on a compute node with 192 GB RAM (the minimum required for the pipeline is 192 GB) and two Intel Xeon Gold 6130 CPU that have 16 cores each and run at 2.1 GHz. Each observation requires 2-3TB of storage for input data and intermediate products. The surveys project makes continuous use of 10-20 of such nodes spread between Hertfordshire (the LOFAR-UK cluster), Leiden, Bologna, Hamburg and SURFsara. Once the pipeline is complete, the data, final calibration solutions and images are uploaded to a storage cluster in Leiden. These final DD legacy products for each pointing occupy approximately 250GB and the total volume required to store the entire survey will be approximately 800TB. The pipeline used for this step is described in Shimwell et al. (2019) and Tasse et al. in prep.

Technical processing solution
The LOFAR archival data uses distributed storage that the SKSP accesses via X.509 certificate-based federated AAI. The compute clusters at SURFsara and PSNC are also accessed via these certificates whereas the other compute cluster require individual ssh key-based solutions.

The SKSP workflow is based on containerized (*Singularity*) software that is built using continuous integration via *SingularityHub* and *Github* and deployed via the *Softdrive* virtual drive. Workflow orchestration and monitoring is done via the *AGLOW* package, consisting of the *Grid_LRT* framework (Mechev et al. 2017a) and an Airflow-based workflow engine (Mechev et al. 2018b). To optimize the workflow the *Pipeline collector* profiling package (Mechev et al. 2018a) runs concurrently with the pipeline at less than 1% of the total CPU cost. To date ***7 Petabyte*** of surveys data have been processed using this solution and made available via web-based repositories at SURFsara and Hertfortshire.

Future perspectives & SKA1-Low: SKA1-Low will provide 3 times the bandwidth of LOFAR and about 9 times the number of stations. Simple scaling implies that size of the data products will increase by a factor 200 relative to the numbers mentioned above. i.e. the DI and DD products will each move from 250GB to 50TB. Note that DYSCO compression will be able to reduce this number by a factor 3-4 and hence whilst these products are large it is not impossible to envisage moving them between system with sufficiently fast data connections.

The automation and workflow orchestration of the current LOFAR Surveys processing pipelines rely massively on the underlying Grid infrastructure and associated tooling. This Grid model will need to change in the coming years to accommodate the requirements of the high-luminosity LHC and the SKA. These changes are a topic of vigorous discussion and we are now faced with the daunting task of shaping the new environment for these instruments. However, the need for this new eco-system also provides an opportunity for the astronomical community to re-think their solutions and make use of the latest technologies to, (i) integrate compute and storage models from the edge to centralized HPC/Cloud facilities and data lakes, (ii) harvest and deploy software using continuous integration and continuous delivery standards in collaboration with the community. This will undoubtedly change the way in which (radio)astronomy is carried out in the next decade.

In the SKA-era not only the data, but also the software and the users must move to the centralized compute facilities and these facilities must be able to cope with the diverse needs of the radio-astronomical community. The data sizes, organization and complexity of radio astronomy observations with LOFAR and the future SKA (100x LOFAR) also break open a new realm in compute and workflow requirements. Whereas high energy physics (HEP) case consists of a limited set of instruments with many small event files, the SKA case will provide a large set of observing modes and thus a very diverse output in terms data types, sets and sizes.

The SKA imaging case, as prototyped by the LOFAR Surveys, shows that the radio astronomy data cannot be treated fully independently and consists of complex workflows with diverging and converging data flows. Orchestration and (auto)-tuning, possibly assisted by machine learning, of these workflows across different infrastructures (geographically, platform & hardware) poses one of the greatest challenges for the SKA. In addition, further optimization of the underlying algorithms to become fit for purpose in 2025 is another challenge. As an example, we will here mention the imaging task of the pipeline that in the coming years will moving from 20k x 20k pixels to 100k x 100k pixels. If this task is performed will current algorithms the RAM memory requirements will increase from 192 Gigabyte to more than 2 Terabytes of memory.

Breakdown of resources used across the three stages listed above:

- CPU corehours (unit: hrs) -> 1:2:3 = 300:1000:5000
- RAM peak memory (unit: GB) -> 1:2:3 = 16:64:192
- Scratch storage (unit GB) -> 1:2:3 = 400:80:3000

This clearly shows that part-3 the most intensive. Here the CPU corehours correspond to the total of all parallel jobs belonging to one of the three stages, whereas as the RAM and scratch are the amount per job (for the dominant task). It is important to note that the parallisation of the parts is different. For SKSP processing, stage-1 consists of 488 jobs, stage-2 of 982 jobs and stage-3 of a single job. A breakdown of the processing in part-3 is shown in Figure 3.

The pie chart below shows each of these different steps and the time taken for **Part 3**, the direction dependent calibration and imaging stage. The processing starts at the top of the pie chart and to complete the entire process takes 5-7 days (the hatched region shows 1 day). The white regions are imaging and the dark regions are calibration with kMS and DDFacet respectively. The black regions at the end are the creation of auxiliary products (things like QU cubes and low resolution images) once the calibration is complete.

The inner bit of the pie chart gives the total breakdown of imaging (white), calibration (grey) and auxiliary (black).

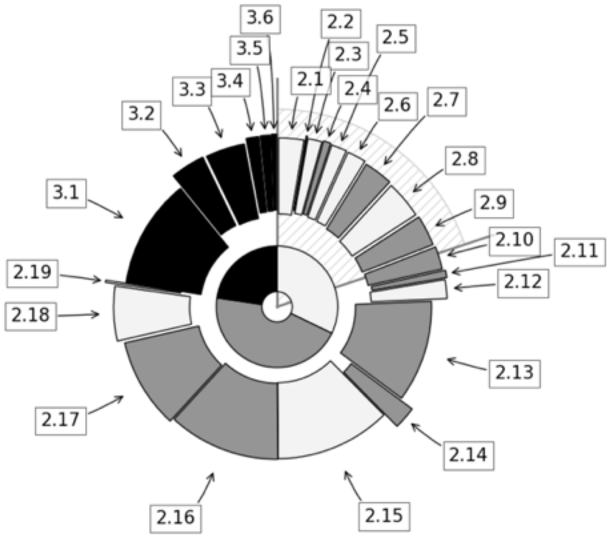


Figure 3: The pie chart below shows the different steps and the time taken for **Part 3** of the surveys workflow, the direction dependent calibration and imaging stage. The processing starts at the top of the pie chart and to complete the entire process takes ~5-7 days (the hatched region shows 1 day). The white regions are imaging and the grey regions are calibration with kMS and DDFacet respectively. The black regions at the end are the creation of auxiliary products (e.g. QU cubes, low resolution images and dynamic spectra) once the calibration is complete. The inner part of the pie chart gives the total breakdown of imaging (white), calibration (grey) and auxiliary (black).

References

de Gasperin F., et al., 2019, A&A, 622, A5
Heesen V., 2018, MNRAS, 474, 5049
Mechev A., et al., 2017, Proceedings for International Symposium on Grids & Clouds, 2
Mechev A., et al., 2018a, A&C, 24, 117
Mechev A., et al. 2018b, arXiv, arXiv 1808.10735 (IEEE)
Offringa A. R., 2016, A&A, 595, A99
Shimwell T. W., et al., 2017, A&A, 598, A104
Shimwell T. W., et al., 2019, A&A, 622, A1
Tasse C., 2014, arXiv, arXiv:1410.8706
Tasse C., et al., 2018, A&A, 611, 87
Tasse C., et al., in prep
Williams W. L., et al., 2016, MNRAS, 460, 2385
van Weeren R. J., et al., 2016, ApJS, 223, 2

Case 3: Epoch of Reionisation

LOFAR, one of SKA's official pathfinders, has a similar station size, field of view, frequency and baseline-coverage as SKA-Low. The LOFAR-EoR Key Science Project (PI: Koopmans) data-processing pipeline (Figure 4) operates on a significant (132xNVIDIA-K40) GPU cluster and is partly distributed, operating at nearly 100% capacity. *Hence we deem the operational LOFAR-EoR processing pipeline as an excellent , but still limited, starting point for an SKA-Low simulation, processing and analysis pipeline*.

Below is a brief description of the modules implemented in the current LOFAR EoR processing pipeline shown in Figure 4.

Direction Dependent Calibration (Stage 4) take >95% of all resources for LOFAR, so is by far the most intense. For example imaging can be done in a few minutes, but calibration takes days. For SKA-low Stage 4 might take 99%+ of all resources. In order to have a major impact the focus should be on fully distributed near real-time calibration (e.g. on GPU-like architectures).

1) Radio-Frequency Interference (RFI): The first module in data processing is to excise ('flag') bad data, e.g., from man-made RFI. We use the AOFlagger module, based on signal-processing and morphological algorithms, to excise input RFI signals from the simulated SKA-Low data.

2) Data Averaging: Very high time/frequency resolution data, from the correlator, is averaged to lower resolution to obtain a manageable data volume, using the NDPPP module

3) Direction-Independent (DI) Effects: Instrumental (amplifier, cable-reflection) errors, after combining signals from receivers in a station ('beam forming'), and large scale ionospheric errors, are direction independent, meaning that each source in the sky is affected by it in the same manner.

4) Direction-Dependent (DD) Effects: Direction-dependent errors are different for sources seen in different directions on the sky. They can be due to amplifier gain errors in the receivers *before* beam-forming (see step 3), malfunctioning receivers, and the ionosphere.

5) Wide-Field Imaging: Low-frequency imaging is 'all-sky' imaging. We use the modules WSClean on CPUs and ExCon on GPUs, developed by Offringa and Yatawatta, respectively, to generate extremely wide-field data cubes (>30x30 degrees or >10^9 pixels per channel) over the full frequency range. These images form the input of the sky model building module (Step 6).

6) Sky Model Building: Both DI and DD calibration of the instrument and ionosphere (steps 3 and 4) require a precise and accurate 'sky model' (i.e. a set of parameterized source models with their frequency dependence). This sky model is built from the image cubes generated in step 5. The current sky-model building module is labor intensive and uses different codes (i.e. Compact source: BuildSky; Extended sources via 'Shapelets'; Diffuse Emission via Spherical Harmonics).

7) Sky-Model Subtraction: Once the iterative procedure of calibration-imaging has converged, the sky model (compact, extended and diffuse emission) can be subtracted from the data after applying the DD instrumental and ionospheric solutions using SageCal-CO.

8) 21-cm Power-Spectra/Tomography: After step 7 in principle only the (spectrally-varying) 21-cm and noise signals should remain in the residual data-set. We currently obtain power-spectra using a model that performs a Bayesian Gaussian Procession Regression analysis in the frequency direction, combined with a spatial Spherical Harmonic Analysis.

9) Signal Inference: The final step in the simulated and observed data analysis is to infer the parameters of the 21cm signal model.

EOR project Hardware & Observational Data

- **Data Processing:** The LOFAR EOR KSP team operates two GPU-based data processing clusters ('Dusk' and 'Dawn'), the latter being the most powerful. It has 1584 hyper-threaded cores (24TFLOPS; 4.6TB memory; 1.7PB internal storage; 10GB/s connectivity), distributed over 32 nodes (plus 1 server) each with four K40C NVIDIA GPUs, yielding 0.55/0.18 PFLOPS at single/double precision.

- **Data Acquisition:** Beam-formed station voltages go via glass-fibers to Groningen, where a GPU-correlator 'COBALT' generates data at 2s/3kHz, yielding 50-80TB 'raw' data in runs of 8-16hrs. These are transported to our processing clusters via a fast 10Gb/s connection. About 3100hr of data has been acquired since 2012, covering 115-189MHz.

- **Data Storage**: The 'raw' data are cleaned of interference (RFI) yielding a 3-5% data loss. The data is averaged to 2/10/10s and 12/61/180KHz, respectively. About 5PB of pre-processed data is stored on storage clusters (Groningen and ASTRON) and 5PB 'raw' data at various HPC centers: SARA (NL), Jülich (GE) and Poznan (PL).

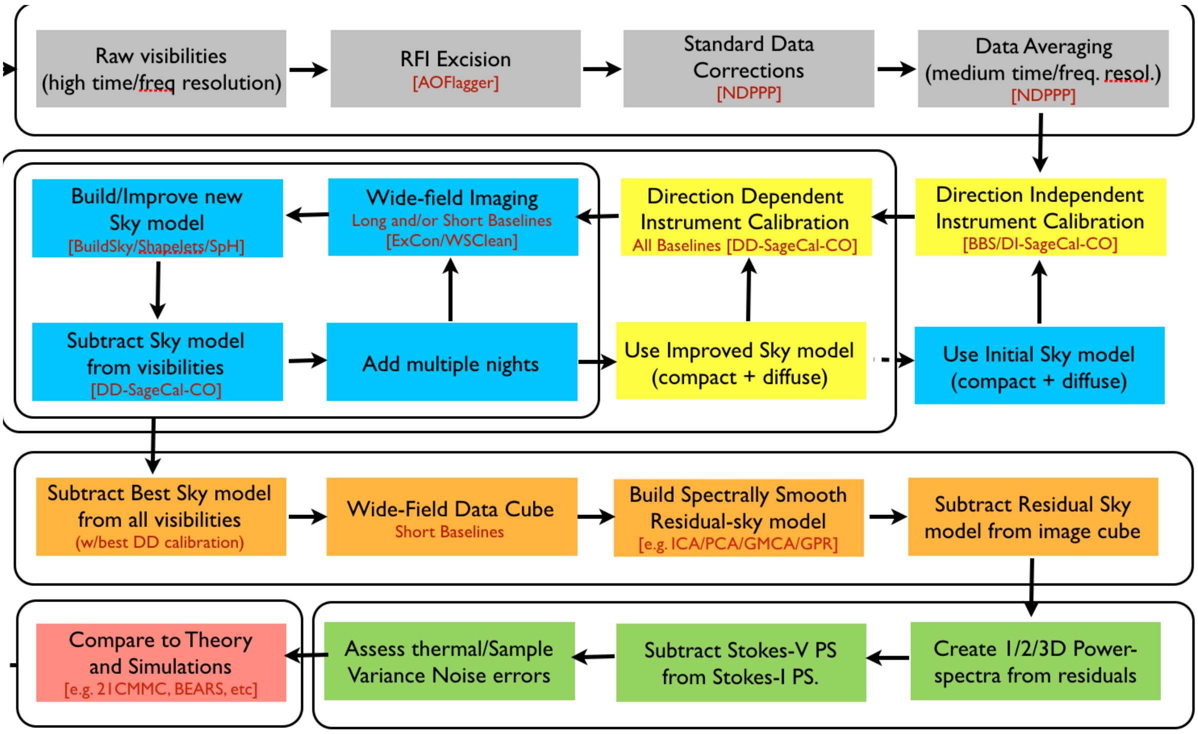


Figure 4: The data processing flow diagram for the LOFAR-EoR pipeline. This forms the basis for the SKA-Low forward simulation and calibration pipeline.