Update from WP3: Computing Requirements

Anna Scaife, Rosie Bolton + WP3 Team

OBJECTIVES

WP3 will identify and assess the components necessary to bring about a European Science Data Centre, both in hardware and software, from a **total science delivery perspective**.

The focal questions are:

"What does the ESDC need to do to maximise European science delivery from the SKA?"

"How can we build such a science data centre and at what cost?"

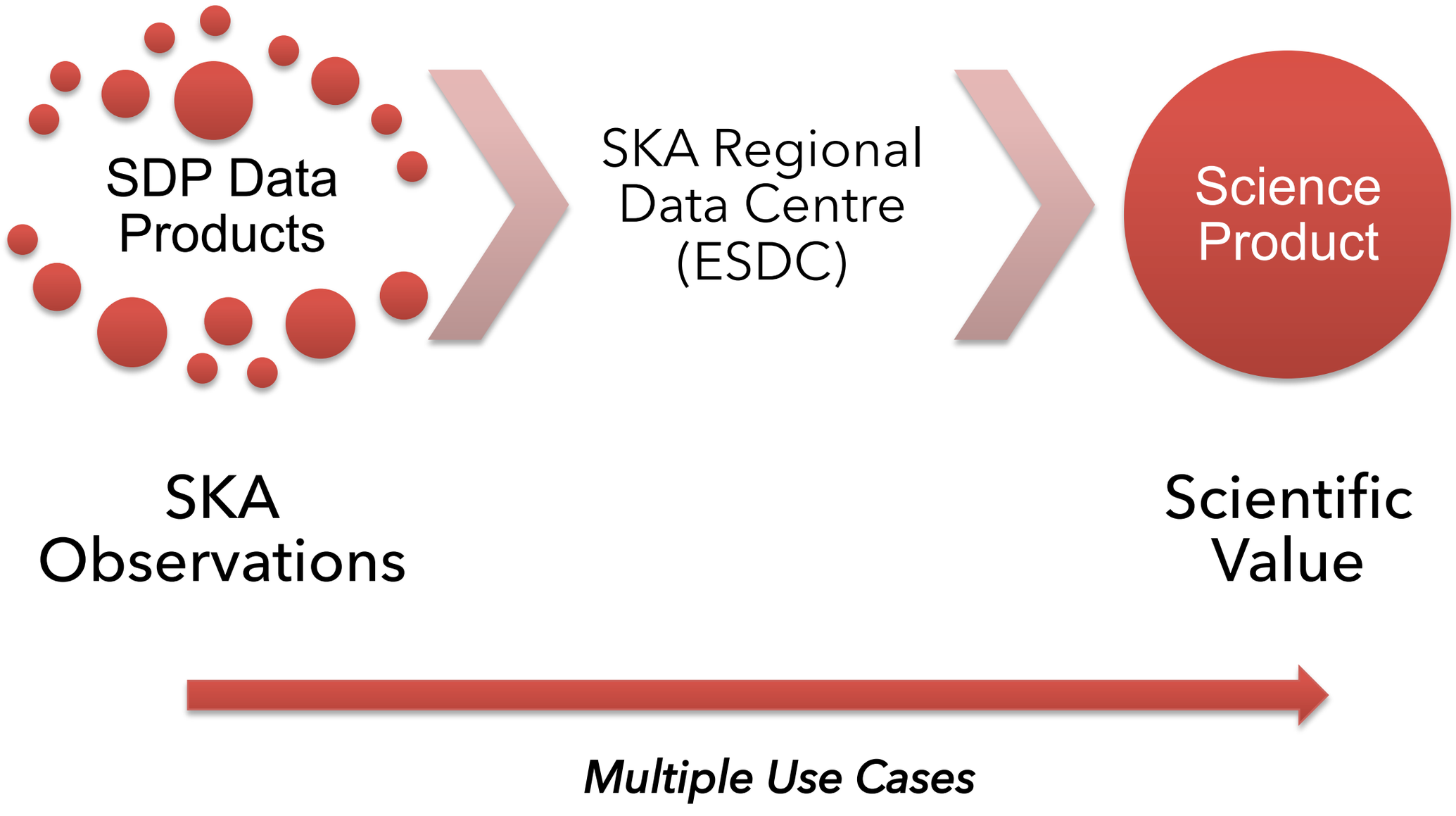
OBJECTIVES

- Develop a set of **design recommendations** for the ESDC pertinent to (1) data handling strategy, (2) scientific functionality and (3) software environment.

- Produce a **high level architectural design** for the ESDC with a **sizing** and **costing** estimate.

- Provide **supporting verification work**, including both theoretical analyses and direct prototyping of critical elements.

- Identify **gaps**, highlight **risks** and make recommendations with respect to **mitigation**.


SDP Data Products
SKA Regional Data Centre (ESDC)
Science Product
SKA Observations
Scientific Value
Multiple Use Cases

WORK BREAKDOWN

T3.1 ESDC Processing: Inventory of SKA science cases and post-SDP computing requirements

T3.2 ESDC Data storage: Inventory and sizing of SKA science data products and ESDC user-derived products

T3.3 Evaluation of existing HPC, cloud and distributed computing technologies

T3.4 Design and costing for distributed ESDC computing architecture

T3.5 Requirements for interfaces to SKA Science Archives & Other Repositories

T3.6 Validation, Verification & Proof of concept activities utilizing SKA pathfinder and pre-cursor facilities

WP3 : Milestone Round-Up

[WP3-lead milestones]

M02	Preliminary functionality assessment		Mark Ashdown
M12	Analysis of compute load, data transfer and data storage anticipated as required for SKA Key science		None
M13	Detailed schedule of anticipated SKA-related data products and their storage requirements		Therese Cantwell
M14	Middleware FoM review		Eva Sciacci
M15	Top-level software FoM review		Mark Ashdown
M16	Full functionality assessment		Rohini Joshi
M17	Test data sets available		Rohini Joshi
M23	Performance required to enable synergistic science incorporating multi-wavelength surveys		Alex Clarke

Draft available/complete

Milestone slightly overdue; no draft available

Milestone very overdue; no draft available

Middleware 1/2

Software products and middleware solutions for allowing access to distributed computing facilities (cloud compute, HTC, HPC and container-based compute).

M14 Document:
https://docs.google.com/document/d/1FIujDiz9Bj73aV7qdpjddWHy7kbw0nIKuEGCuIHq0hY/edit?usp=sharing

- Expected to be a “living” document to be updated during the lifetime of the project.
- Initial inventory of available middleware products and solutions:
  Schedulers and Workload Management; Programming models and frameworks; Containerisation and virtualisation implementations; Cluster management; Data management; Monitoring Tools
- Cloud review integration is in progress (EGI)

Middleware 2/2

- FoMs Definition
  - General FoMs starting from the literature
    - E.g.: Scalability; Usage across supercomputing centers or in other experiments; Active development/Active community/Sustainability; etc
  - Data Management specific FoMs
    - E.g.: Data handling capabilities/data type-size-volume; Efficient movement of data based on policies; Support for metadata description; etc.
  - Schedulers and Workload Management specific FoMs
    - E.g.: Is there support for interactive jobs? MPI jobs? Parallel and/or array jobs?
  - Federation Requirements specific FoMs
    - E.g.: Support for Usage Accounting/Logs; Integration with AAI (e.g. LDAP or others); Security management...
- Waiting for FoMs coming from requirements analysis collected in T3.1 and T3.2 (and also from survey results)
- Rank and evaluation of middleware technologies based on the FoMs

See talk in WP3 session on Monday afternoon

Top-level software stack

Programming framework for the parallel and distributed analysis of data
M15 document:
https://docs.google.com/document/d/1Li_U_LT3uFsqouPoXisRZyp4Rh2OyHKyDcEwysEIgy8/edit?usp=sharing

- Expected to be a “living” document to be updated during the lifetime of the project.
- Brief review of data types and use cases
- Definition of the figures of merit
- Description and evaluation of the programming frameworks:
  - Spark; Dask; Legion; Swift/T; PGAS languages (e.g. Unified Parallel C, Coarray Fortran); MPI + X (where X supports multithreading, e.g. OpenMP)

See talk in WP3 session on Monday afternoon

Test Data

Available via:

- Direct download on Zenodo
- On CVMFS on WLCG
- Image dataset (TGSS - 20k x 20k)

Still to come:

- Time domain data

See talks in WP3 session on Monday afternoon & WP3-WP5 session on Tuesday afternoon.

T3.6 Prototyping Activities

- Implementing example use cases on existing infrastructure
- Mostly using WLCG resources
- Identifying potential bottlenecks in e-infrastructure designed for other fields

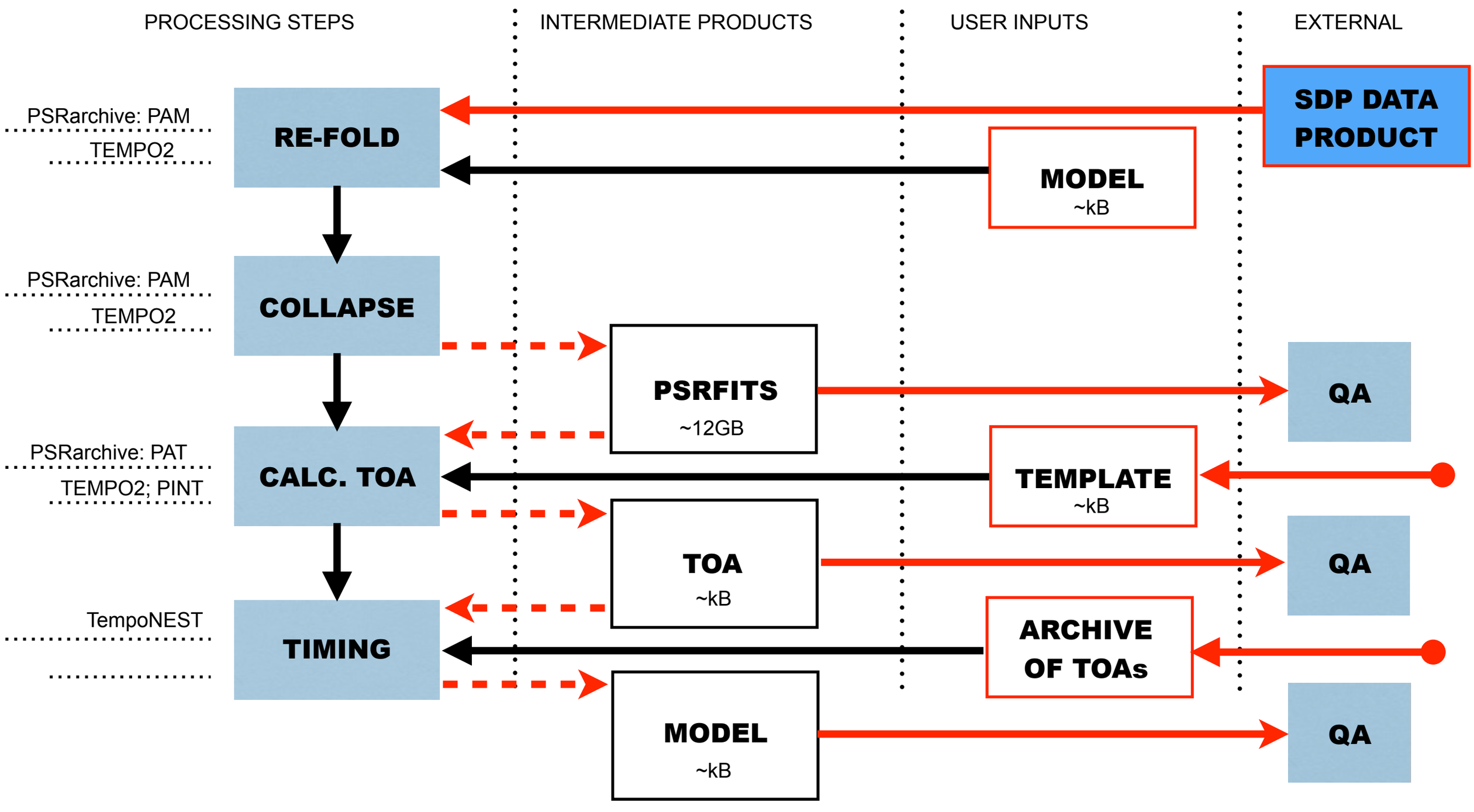
Current Compute Model Use Cases:

- Calibration & Imaging Use Case
- Image-based Object Detection & Classification Use Case
- Catalogue-based Cross Matching incorporating External Archives Use Case

- Image Mosaicking Use Case
- Image Cube Stacking Use Case
- Time-domain Re-folding Use Case

See talks in WP3 session on Monday afternoon

COMPUTE MODEL USE CASE : Transients SWG



OBJECTIVES

- Develop a set of **design recommendations** for the ESDC pertinent to (1) data handling strategy, (2) scientific functionality and (3) software environment.

- Produce a **high level architectural design** for the ESDC with a **sizing** and **costing** estimate.

- Provide **supporting verification work**, including both theoretical analyses and direct prototyping of critical elements.

- Identify **gaps**, highlight **risks** and make recommendations with respect to **mitigation**.