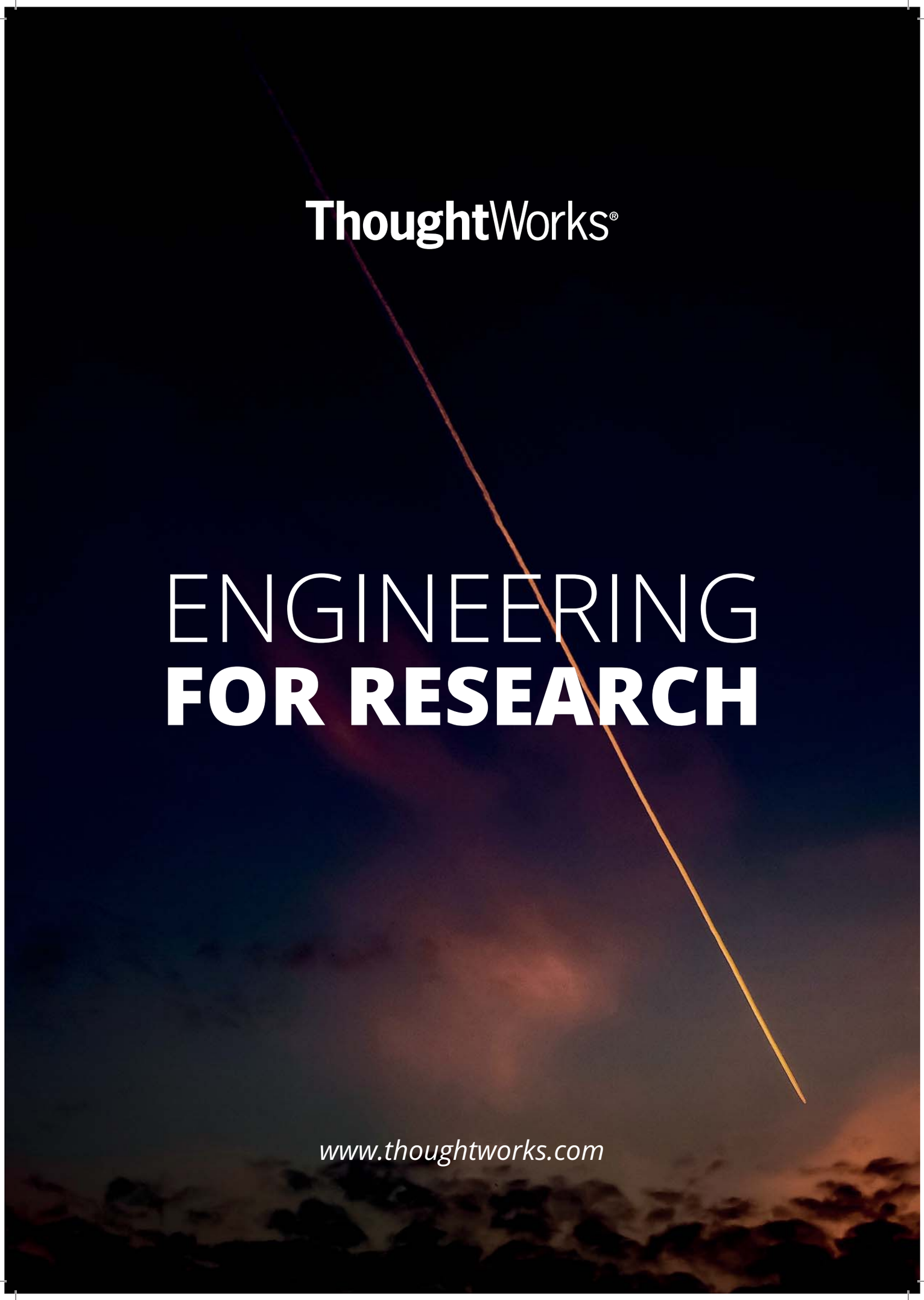
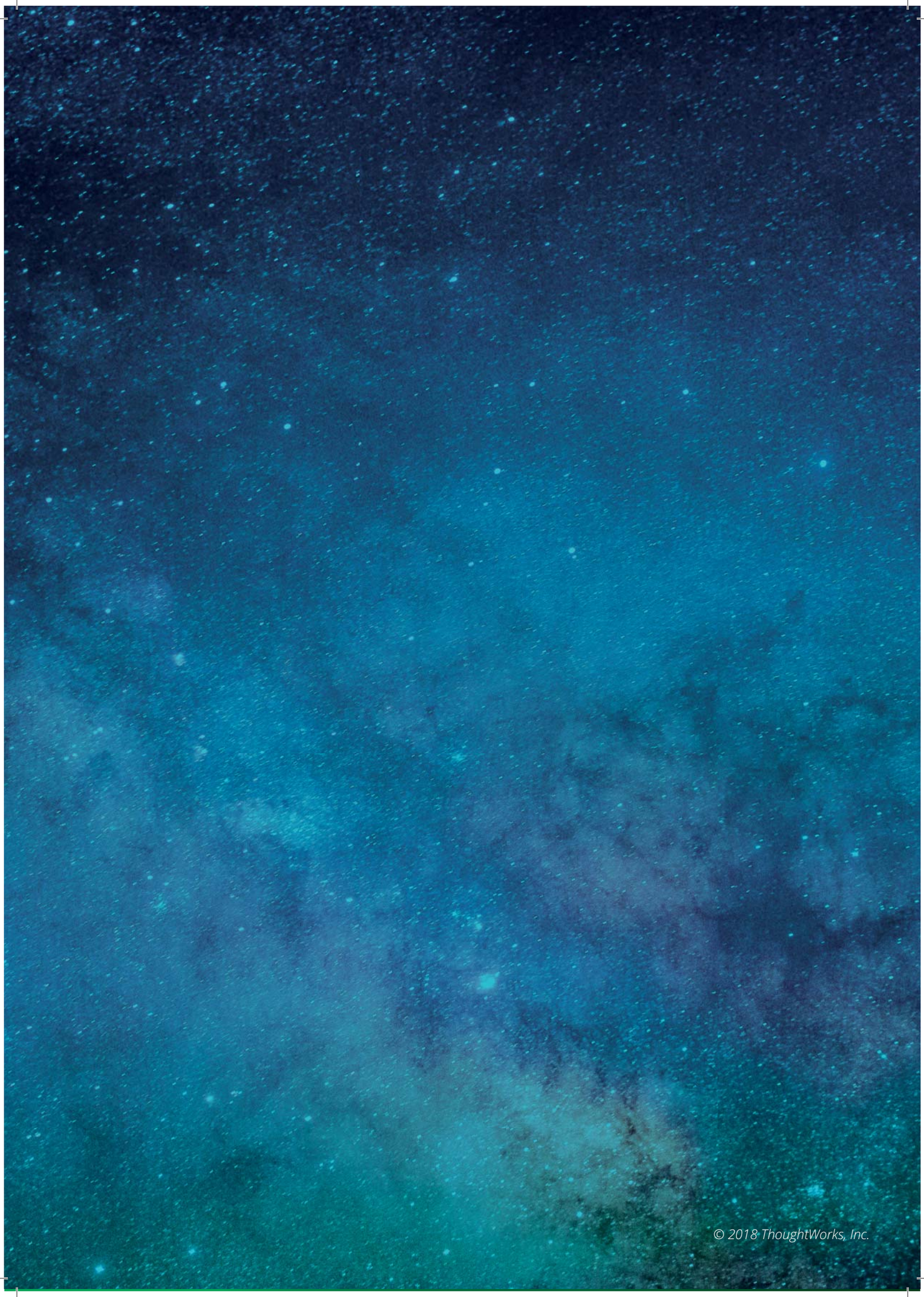

ThoughtWorks®
ENGINEERING
FOR RESEARCH
www.thoughtworks.com



OVERVIEW

ThoughtWorks Engineering for Research or E4R is our initiative to advance research in the fields of radio and optical astronomy, genomics, molecular dynamics, and urban sciences by working with leading scientific organizations from around the world.

Our work with astrophysicists and scientists has brought into focus the ever-increasing volumes of data within scientific disciplines or the **Data Deluge** that is causing a shift in how science is evolving. This shift is otherwise referred to as **Data-intensive Scientific Discovery** or **The Fourth Paradigm** and calls for a rethinking of current computational tools, frameworks and processes.

Armed with experience that spans global collaborations across both, domains and technologies, ThoughtWorks is primed and committed to tackle the data challenge in the area of scientific research. The plan is to engage in broader, deeper and long-term engagement with the scientific community that is aimed at discovering new approaches, tools and frameworks using everything from Artificial Intelligence to Machine Learning to Cyber-physical Systems to Data-intensive Computing to Digital Libraries, Lab Informatics and Simulations.

CLIENT STORIES

THIRTY METER TELESCOPE

Engineers, astronomers, and project specialists are working together to build the highly anticipated, next-generation observatory for the astronomical community, the **Thirty Meter Telescope (TMT)**. With TMT, we will be able to study the universe as never before, finding answers to many of the grand challenges of science such as the origin of galaxies, birth and death of stars, probe turbulent regions surrounding black holes, discover planets orbiting distant stars and the possibility of life in the alien worlds.

ThoughtWorks in association with **Indian Institute of Astrophysics** is collaborating with the TMT team to build the following systems:

1. **Common Software (CSW)** are part of Level 2 TMT subsystems that provide TMT software technical architecture and infrastructure, required to integrate all of TMT software.

2. **Data Management System (DMS)** is the system that provides the software and hardware mechanisms, and infrastructure to capture, time-stamp, describe, store, transmit and access all science and engineering data flowing through the TMT system.

3. **Executive Software System (ESW)** provides core functionality for synchronized operation of all TMT subsystems apart from observing and monitoring user interfaces.

    a. **The Observatory Control System (OCS):** The centre that implements data sequencing to support the observation data acquisition, and shares the execution information across systems.

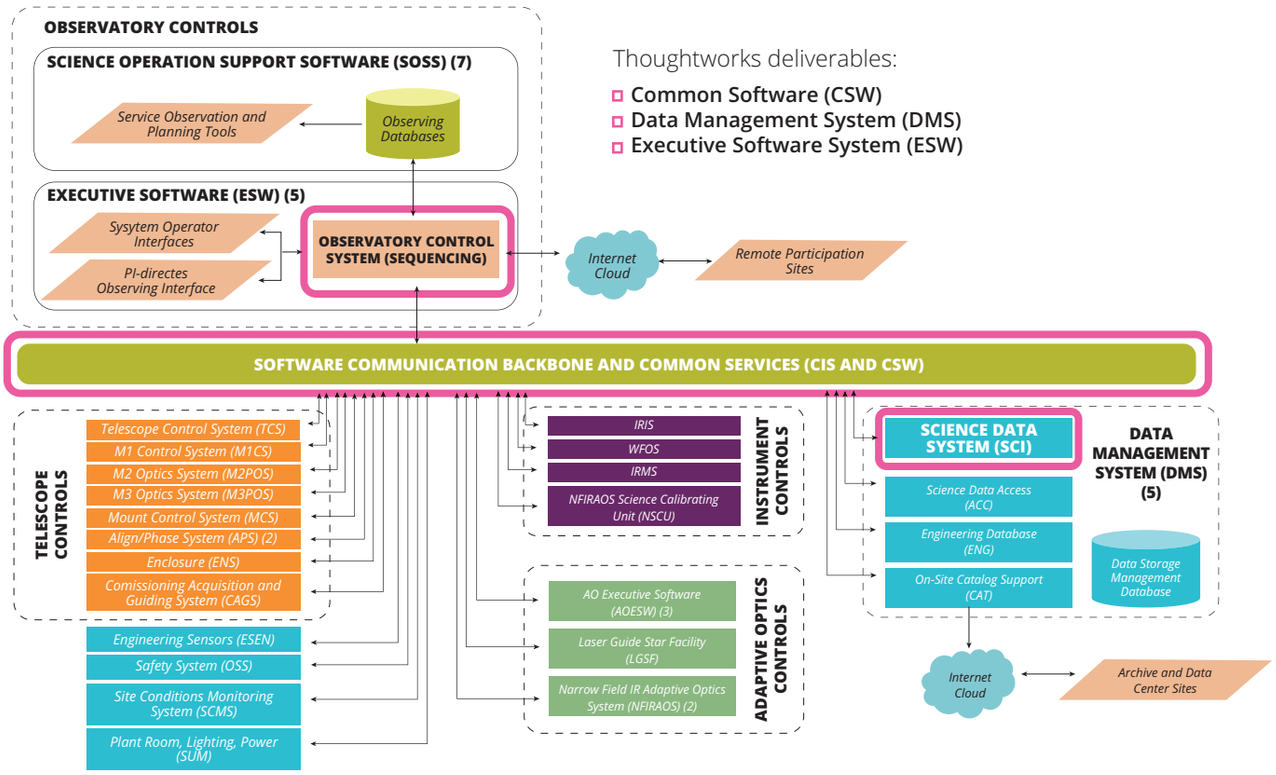
The work done for building the location service in the Common Software system was presented as a session on Service Discovery using CRDT at **React Summit 2017**, Austin, TX, USA.

CLIENT TESTIMONIAL

The TMT Project Office Observatory Software team is very pleased with ThoughtWorks' contributions to the Common Software development project. The ThoughtWorks team has been very enthusiastic and engaged throughout the project; they communicate frequently, clearly and transparently. The team applies its significant experience and knowledge to develop optimized design solutions and robust implementations that meet TMT's functional and performance requirements. They have implemented continuous integration and automated testing pipelines to meet TMT's rigorous QA requirements. We meet with them over video-conferencing at least twice a week, and every 4 – 6 months we meet face-to-face to plan the next development phase; the meetings are always friendly, respectful and productive. It is a true pleasure to work with the ThoughtWorks team, and we look forward to a long-term relationship focused on the development of the many TMT Observatory Software Subsystems.

- Hanne Buur, Observatory Software Project Manager,
TMT International Observatory

TMT SOFTWARE ARCHITECTURE



TECH STACK:

LANGUAGES	FRAMEWORKS	CI / CD	TESTING	DOCUMENTATION	MESSAGE TRANSPORT PROTOCOLS	STATIC CODE ANALYSIS TOOLS
Scala	Akka Clustering	AWS	Multi-jvm	Paradox	Protobuff	Scalafmt
Java	Akka Streams	Jenkins	Multi-node	Github pages	Json	Scalastyle
	Akka Http	Ansible	wrk	Swagger	Kryo	SCoverage
	Akka Actors	Bintray	postman			
	svnkit					

TMT COMMON SOFTWARE INTEGRATION FRAMEWORK

The TMT Software System is a distributed software system. Software within TMT subsystems is built using a TMT-specific software development framework called **Common Software Integration Framework**, a deliverable of ThoughtWorks. The integration framework provides software that assists in proper use of the services and software that assists in providing best practices for structuring components.

From a software communications and integration viewpoint, TMT consists of a set of software components interacting with each other through a communications backbone and software infrastructure (middleware). The middleware infrastructure is a collection of various channels corresponding to different command, configuration, or data transfer services required for the integration task. The connections shown in the architecture diagram (black lines between TMT subsystems) correspond to one or more such channels.

TMT COMMON SOFTWARE SERVICES:

LOCATION SERVICE
Locate, register and track changes for a component's connection information

COMMANDS SERVICE
Support receiving, sending and completion of commands in the form of configurations

LOGGING SERVICE
View, capture and store local and distributed logging information

COMMON SOFTWARE (CSW) INTEGRATION FRAMEWORK
Templates for various software components defined by TMT as well as service access interfaces

CONFIGURATION SERVICE
A centralized persistent store for any configuration file used in the TMT Software System, with versioning system to provide a historical record of each configuration file.

AUTHENTICATION AND AUTHORIZATION SERVICE
Centrally manage user authentication/access control

TELEMETRY SERVICE
Publish/subscribe system for component status and telemetry

ALARM SERVICE
Support component alarms, and component and subsystem health

TIME SERVICE
Standards-based, precision time access for synchronization

EVENT SERVICE
Publish/subscribe system for demands and other transient events

DATABASE SERVICE
Access to a shared, centralized, relational database

INTER-UNIVERSITY CENTER FOR ASTRONOMY AND ASTROPHYSICS (IUCAA)

Next generation telescopes like **MeerKAT and Square Kilometer Array (SKA),** generate huge amounts of data. The SKA, for instance, when completed by the mid 2020's is expected to produce more data in a day than the entire internet!

The MeerKAT radio telescope, precursor to the SKA, is currently under construction in the Karoo desert of South Africa. Until SKA is completed, MeerKAT will be the most sensitive telescope at cm wavelengths and will propel transformational science ahead of SKA.

Dr.Neeraj Gupta, Associate Professor at IUCAA in Pune, has approximately 1700 hours of observing time at MeerKAT to carry out a large survey, the MeerKAT Absorption Line Survey *(http://mals.iucaa.in/)* or MALS which will map the evolution of cold gas in the universe. As processing the data *(approx. 4 PB over 5 years)* from this survey using conventional methods, that also involve manual interventions will take decades - automating the data processing techniques has become the need of the hour.

To address this, ThoughtWorks and IUCAA are developing an **Automated Radio Telescope Image Processing Pipeline (ARTIP)** that will automate the entire process of flagging, calibrating and imaging while processing the data. The pipeline will also use various statistical techniques to identify bad data patterns like completely or partially bad antennas, baselines and timeranges, apart from generating flagging statistics and various diagnostics at each stage of the pipeline.

ARTIP has been tested and validated against various datasets (each size: approx. 10GB) from the Giant Meter Wave Radio Telescope (GMRT) and the Very Large Array (VLA) telescopes. The time taken to run the pipeline sequentially, on a server class machine with 256GB RAM and 40 cores is around **20 minutes,** as opposed to a manual process that would take anywhere between **3 to 4 hours.**

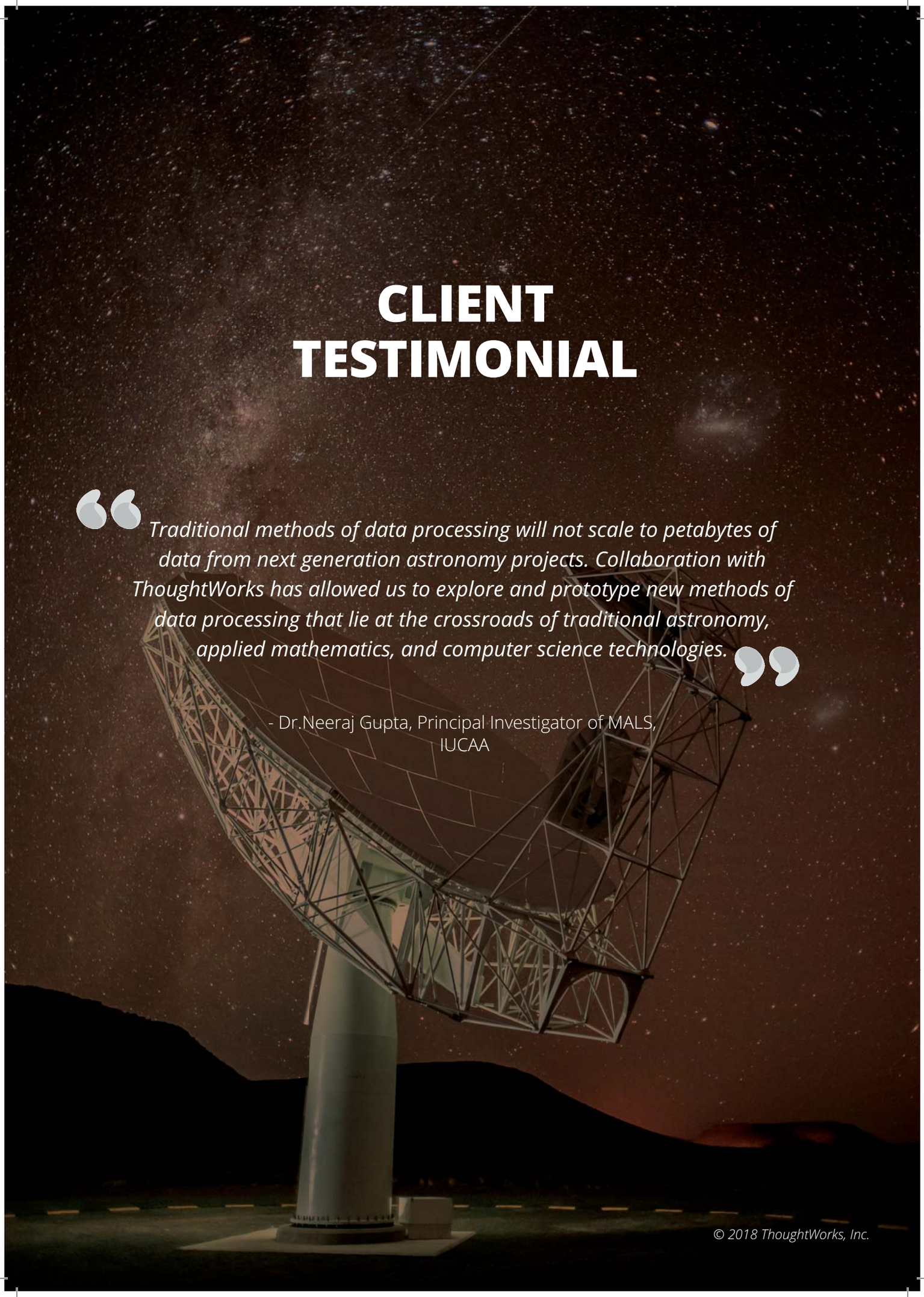
The pipeline runs sequentially, and does not utilize hardware at full capacity. Further enhancements have been planned to scale the pipeline to handle petabytes of data. For parallelization, the team is currently working on the imaging stage, which is takes more than 70% of the total processing time. The approach is to partition the data in frequency axis and run imaging parallely on a 13 node cluster (each node having 128GB RAM and 16 cores) provided by IUCAA, backed by a Lustre File System. This gives a gain of 60% on a 1.4TB dataset, but the scaling is restricted by the number of logical partitions (useful for achieving science goals) that can be made.

Amongst the key achievements of the ARTIP are:

1. The detection of the OH radical in a galaxy. Such detections are rare (only 3-4 are known so far) and would have been easily missed without the pipeline.

2. Presented at **hiAbsorption 2017**, international astronomy conference, at **ASTRON**, Netherlands.

3. Work for MALS data processing with ARTIP, has been recognized in international astronomical journal, **Proceedings of Science**, in the publication *The MeerKAT Absorption Line Survey (MALS).*

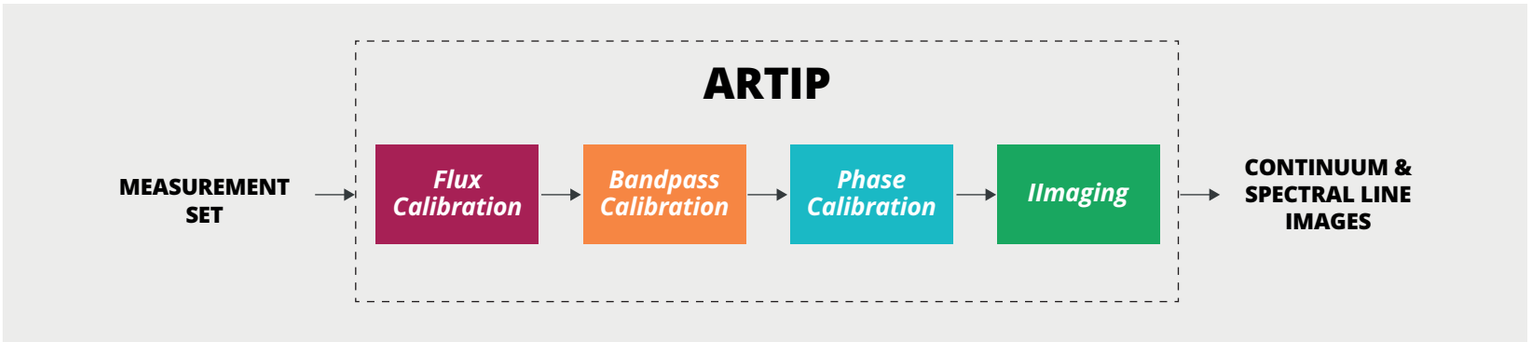


CLIENT TESTIMONIAL

Traditional methods of data processing will not scale to petabytes of data from next generation astronomy projects. Collaboration with ThoughtWorks has allowed us to explore and prototype new methods of data processing that lie at the crossroads of traditional astronomy, applied mathematics, and computer science technologies.

- Dr.Neeraj Gupta, Principal Investigator of MALS, IUCAA

ARTIP IMAGING PIPELINE



ARTIP operates on measurement sets and runs through several stages to generate continuum and spectral line images

TECH STACK:

CASA Common Astronomy Software Applications

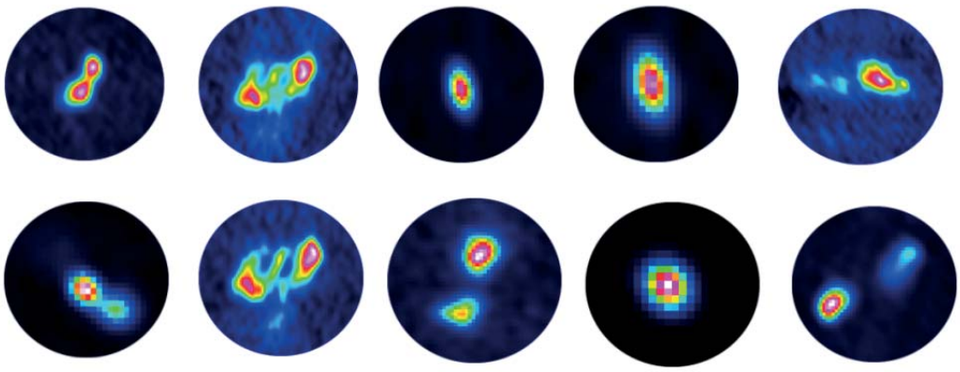
InfluxDB

collectd

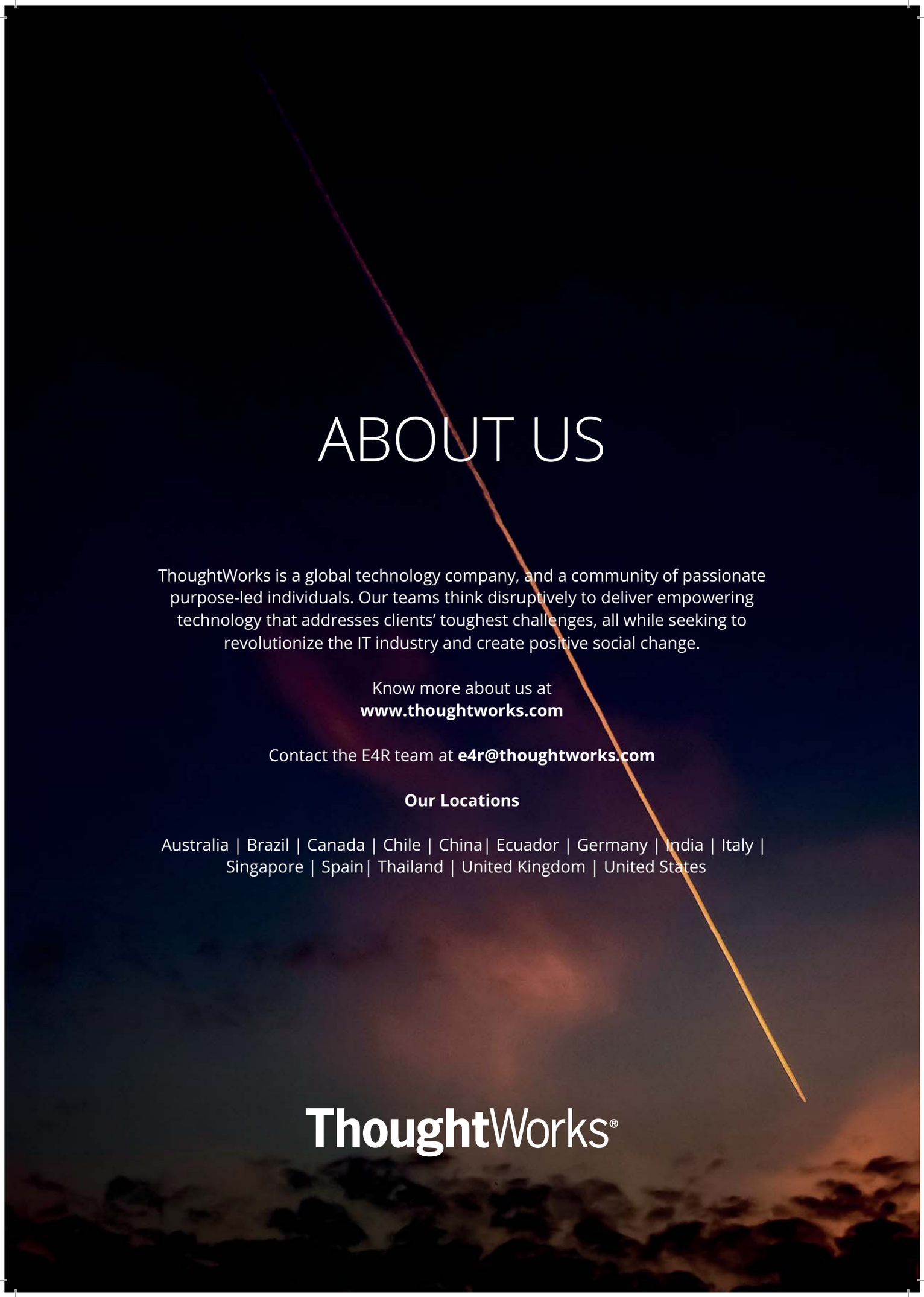
ANACONDA

C3.js

ARTIP IMAGING:



Images of radio galaxies generated by the pipeline from GMRT observations.



ABOUT US

ThoughtWorks is a global technology company, and a community of passionate purpose-led individuals. Our teams think disruptively to deliver empowering technology that addresses clients' toughest challenges, all while seeking to revolutionize the IT industry and create positive social change.

Know more about us at
www.thoughtworks.com

Contact the E4R team at **e4r@thoughtworks.com**

Our Locations

Australia | Brazil | Canada | Chile | China| Ecuador | Germany | India | Italy | Singapore | Spain| Thailand | United Kingdom | United States

ThoughtWorks®