Big Data architecture in radio astronomy: the SKA

Piyush Nahata
Data Science
SVKM's NMIMS
Nashik, India
piyush1333@gmail.com

Karan Mehta
Data Science
SVKM's NMIMS
Mumbai, India
karanpmehta5161@gmail.com

Prof. Sarada Samantaray
Data Science
SVKM's NMIMS
Mumbai, India
sarada.samantaray@nmims.edu

***Abstract*— The SKA i.e., Square Kilometer Array is a radio telescope project which was conceived in 1990, designed in 2010, implementation of the project started in 2020 along with pilot testing and is estimated to finish its first phase in the year 2027. It is a radio telescope project which will operate to capture radio signal in a very wide range of frequencies. This SKA project will observe the sky and collect data 50 times more sensitive than any other existing radio telescope project in the world and hence expected to be very informative. The plan is it will be operational for the next five decades. To manage this data in a sophisticated form there is a need for an architecture that will help manage the data and process it before it reaches the end-user. Also, it is realized that the amount of data will be so huge, we may not be able to store it for batch processing. In this paper, we are trying to recommend an architecture named Lambda Architecture which might help in faster data flow and overcome the demerits of the old streaming model.**

Keywords—SKA, Big Data, Radio telescope, Lambda Architecture, Radio Astronomy.

I. Introduction

Astronomy is an important approach for us to know the insights of the solar system and for that, we need a radio telescope to get the micro details of the whole Universe, the SKA is one of the radio telescopes which estimates to generate huge data that need to be organized and processed to find insights. SKA is the future of telescopes that can give the micro details of the Universe. The SKA will be one of the telescopes with an aperture of up to a million square meters This project is the biggest milestone of Space Research of the 21st Century. Big data analysis influences space exploration by enabling us to better understand space data and unlock the mysteries of the universe. Since the estimated data generation of this radio Telescope is around 700 Terabytes of data per sec, which will be generated by using around 3000 dishes and 1 million antennas. Once the SKA is launched the data generated will be 10 times more than the global internet traffic today. SKA ultimately produces images and other data at wavelengths from 4.3 m (70 MHz) to 1 cm (30 GHz).

The SKA should be placed as radio-quiet as possible to avoid interference from other high-frequency signals that could contaminate the data produced by the telescope. As such, SKA is widespread in the desert fringes of South Africa and Australia, with international headquarters at Jodrell Bank, UK. 200 mid-range courts (including MEERKAT's existing one) will be built in the Karoo region of South Africa. There are about 16 countries including India in core SKA development and over 100 organizations from over 20 countries participate in the development and formation of SKA.

The implementation of the Square Kilometer Array (SKA) is divided into two Phase:

1. SKA_MID: This Phase1 implementation is started in South Africa in the Karoo region, which is linked with MeerKat by using its dishes. It will give a detailed observation for mid and high frequencies (350 MHz to 14 GHz).

2. SKA_LOW: This Phase2 implementation is started in Australia for this phase of the SKA will be using SKA's dishes and the antenna itself. It will give a detailed observation for the low frequencies (range: 50 to 350 MHz).

The difference of both the phases are shown in the below table1:

Description	Australia SKA_LOW	South Africa SKA_MID
Type of Sensor	256 dipoles each for 512 stations	Including 64 MeerKAT - 197 Dishes
Range(Frequency)	50-350 MHz	0.45-15 GHz
Collecting Area	0.4 Km^2	32 km^2
Max Baseline	65 Km (between Station)	150 Km
Raw Data Output	157 Tb/sec (0.49 Zettabytes/year)	3.9 Tb/sec (122 Exabyte/year)
Science Archive	0.4 PB/day (128 PB/year)	3 PB/day (1.1 Exabyte/year)

Table 1

II. Data Flow of SKA

The Below data flow diagram represents the amount of data at every node until it reaches the end-user. This is the flow of the data that will be generated by the SKA_MID dishes and SKA_LOW antenna. The transmitting speed for data in SKA_MID will be around 8.8Tb/s and SKA_LOW passes data/signals to PFLOPS at the transmission rate of ~2pb/s which are hatched by antennas. Data from both phases of SKA's are passed on to 50PFLOPS and 100 PFLOPS via fiber optics for further operation. After which this data is being represented to the end-user.

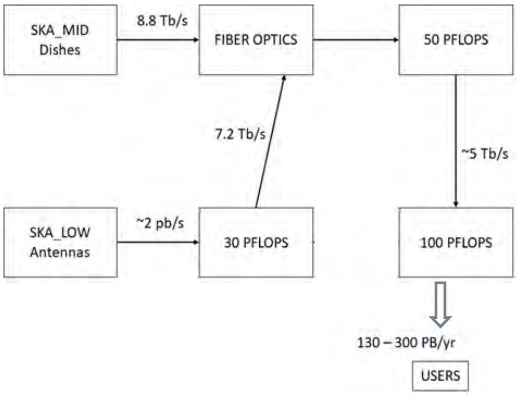

Fig 1. Data Flow from Square kilometer array

The main reason for building this world's most powerful radio telescope is:

- Testing general relativity (strong regime, gravitational waves).
- Cosmic Dawn (First Stars and Galaxies).
- Galaxy Evolution (Normal Galaxies z~2-3).
- Cosmology (Dark Energy, Large Scale Structure).
- Cradle of Life (Planets, Molecules, SETI).
- Cosmic Magnetism (Origin, Evolution).
- Exploration of the Unknown.

Image Quality Comparison:

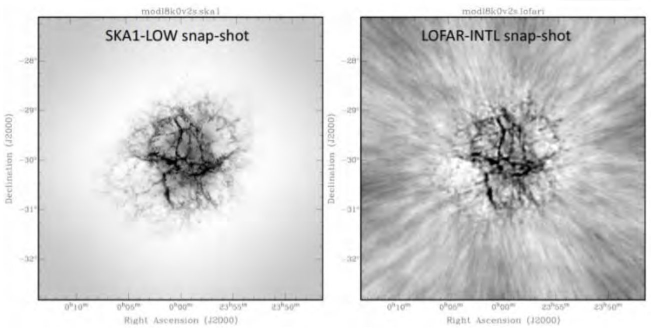

Fig 2. Single SKA1-Low snap-shot compared to LOFAR-INTL snapshot

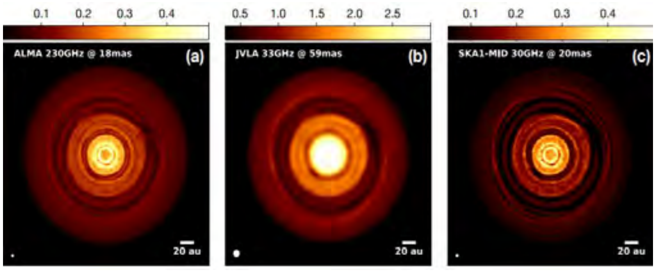

Fig 3. Image of Radiative transfer modelling of a planet from (a) ALMA, (b) JVLA & (c) SKA_MID

III. Methodology

A. Old Architecture: SKA Streaming model

In the Streaming framework, the processing of the data is done at the time when it is created. The streaming process involves multiple tasks on the input stream which can be performed in serial or parallel or both manners. This is called a stream processing pipeline which is the collection of generation, process, and storing to the final location.

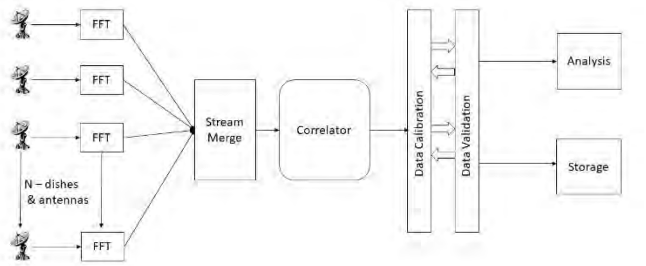

Fig 4. Streaming Model Architecture of SKA

For SKA, when the data is received from the antenna/dishes it is further transformed using Fast Fourier Transform (FFT), which converts the radio signal into representable format then this data is sent to a streaming module where the continuous flow of data is maintained. When the query is fired by an authorized person, the correlator which is located near to the core of Array grabs the data and calculates the cross-correlation function as a function of time lag. This will convert the data is synchronized and combined format. After this filtration and validation of data are done which is then used for image formation or analysis and storage.

The below equation is a translation of the ongoing process of data in the streaming architecture which defines any query in the big data domain.

Query = S (New Data) = S (Live streaming data)

This equation signifies that all queries can be supplied by applying the above function to all the live streams of data at the streaming layer. The Streaming architecture can be built using Apache Kafka.

Demerits of streaming model:

- Without stack layers, errors can occur during data processing or database updates.
- The data is in a multidimensional format which is hard to process while streaming.
- Main memory computation of data.

B. Recommended Model: SKA Lambda Architecture:

Lambda architecture is the combination of both Micro - Batching and Streaming architecture. To deal with huge amounts of data in an efficient manner Lambda Architecture can be used. The efficiency can be measured by looking at the increased throughput, reduced latency and negligible errors in Output Data.

The data generated in a radio telescope is exceptionally high and requires high throughput with minimum latency possible and with fewer errors in output data. This can be achieved by the below architecture which also overcomes the disadvantages of the old architecture i.e. absence of the batching layer. Since the model contain both the batching and streaming layer the errors during data processing or while updating the database is reduced or is negligible

The below equation is a translation of data queries in the big data area. This symbol 'λ' is Known as the Lambda function.

Query = λ (Complete data) = λ (live streaming data) * λ (Stored data)

This equation signifies that all queries related to data can be supplied in Lambda architecture which will combine the latest data in the past results in the form of batches and stream of the data can be done by streaming layer.

The major mmanagement part of this huge flow of data and writing queries will be done with help of big data tools like:

- For batching tools: Hadoop MapReduce, Spark, PIG
- For streaming tools: Apache storm, Apache spark Streaming, Spring XD
- For Batching database tools: DRUID, Elephant DB, HBase
- For Speed Layer database tools: Riak, Cassandra, Redis

The Lambda architecture comprises of following modules:

- Micro - Batching Layer
- Speed Layer or Stream Layer
- Serving Layer

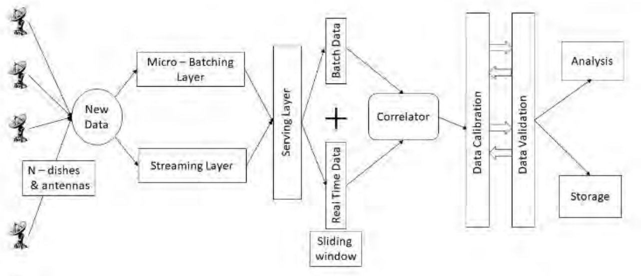

Fig 5. Lambda Architecture of SKA

Micro Batching Model:

Micro Batching framework is the one in which we form a small group of data or we divide the data into batches to process. The old traditional batching process involves the grouping of data in large groups. Macro batching is the intense version of batching where the processing of data is done more frequently to ensure the smaller group of new data is processed using the MapReduce or using ML (Machine Learning) to predict the upcoming batch views.

Speed Layer:

The speed layer takes the advantage of event sourcing done at the Micro Batching Layer. The output of the batch layer which is generated using MapReduce or ML is then used by the stream layer to process the new data given to it. The speed layer generates the output on the basis enrichment process and supports the serving layer to reduce the latency in responding to the queries. Since the speed layer deals with real-time data only and also has less computation load the latency is quite low during the processes.

Serving Layer:

The outputs from the batch layer in the form of batch views and from the speed layer in the form of near-real-time views are forwarded to the serving layer which uses this data to cater to the pending queries on an ad-hoc basis.

Merits and Demerits of Lambda Architecture:

Merits:

- Even if the system crashes, the Lambda architecture is less error prone because the batch layer uses fault-tolerant distributed storage to manage historical data.
- It's a good balance between speed and reliability.
- A scalable and fault-tolerant computing architecture.

Demerits:

- This can lead to coding overhead as it requires extensive processing.
- Reprocess each batch cycle, which is not useful in certain scenarios

IV. Conclusion

The old model i.e., the streaming model has few demerits which can be overcome using the recommended architecture i.e., Lambda since this architecture comprises both the Streaming and Batching layer which makes the model more sophisticated and faster with less error and low latency in data and it will help the data also to flow fast and store the data in an organized form by using big data tools.

V. Reference

[1] Astronomy and Astrophysics in the New Millennium. Washington, D.C.: National Academy Press, 2001.

[2] https://www.skatelescope.org/participating-countries/

[3] SKA Telescope Manager (TM): Status and Architecture Overview: https://core.ac.uk/download/pdf/154355299.pdf

[4] A brief introduction to two data processing architectures — Lambda and Kappa for Big Data https://towardsdatascience.com/a-brief-introduction-to-two-data-processing-architectures-lambda-and-kappa-for-big-data-4f35c28005bb

[5] SKA-Community-Briefing-18Jan2017 https://www.skatelescope.org/wp-content/uploads/2017/02/SKA-Community-Briefing-18Jan2017.pdf

[6] SKA1 Beyond 15GHz: The Science case for Band 6 https://www.skatelescope.org/wp-content/uploads/2020/02/ScienceCase_band6_Feb2020.pdf

[7] Danny Price (2019), Real-time stream processing in radio astronomy https://arxiv.org/abs/1912.09041