Programming Paradigms for Big Data

Surajnath Sidh & Saumya Gupta

Big Data- The Buzzword

Gartner’s definition

The Three V’s of Big Data

- **Volume**: big data doesn't sample; it just observes and tracks what happens
- **Velocity**: big data is often available in real-time
- **Variety:** big data draws from text, images, audio, video; plus it completes missing pieces through data fusion

Some other ideas……

- Unstructured extremely large datasets.
- A cultural movement.
- Big data stands for important data.

What's Behind It?

Facebook- 300 PB data warehouse (600 TB ingested daily)

Google- 40000 search queries per second

Twitter- 500 million tweets per day

Amazon- Estimated about 450000 servers.

Challenge-

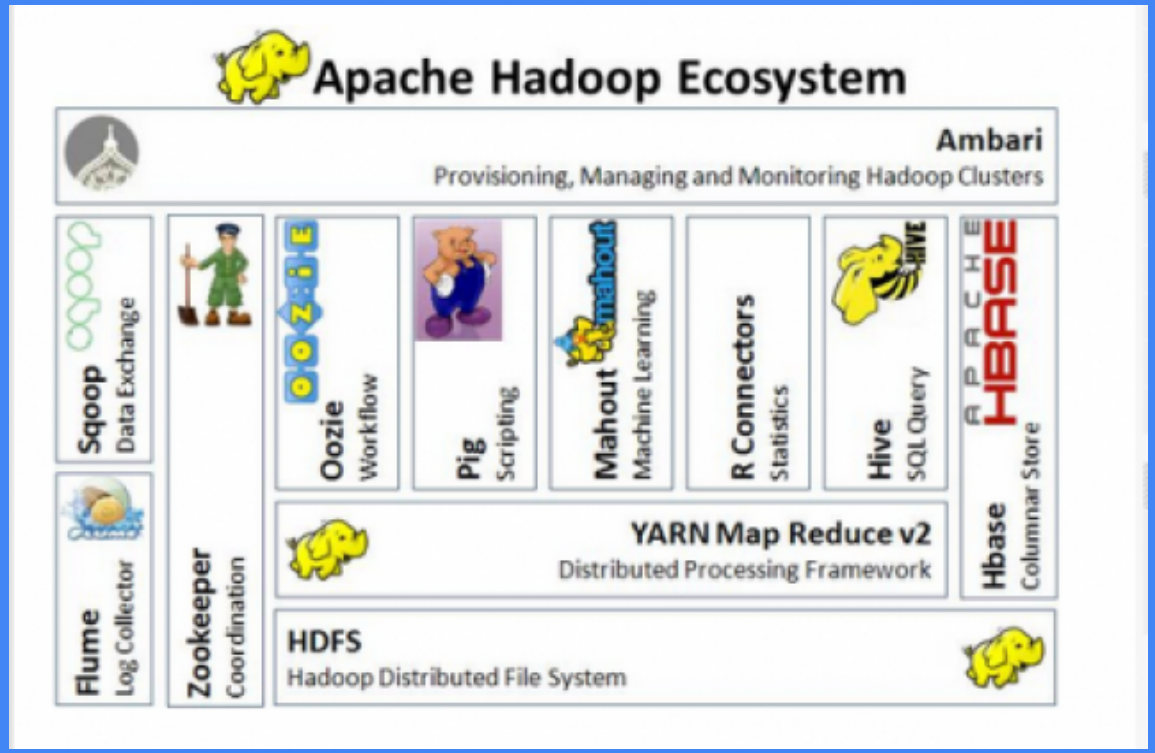
How to process and store data efficiently in this case?

Customer Insight

Motivation

Some Related Work

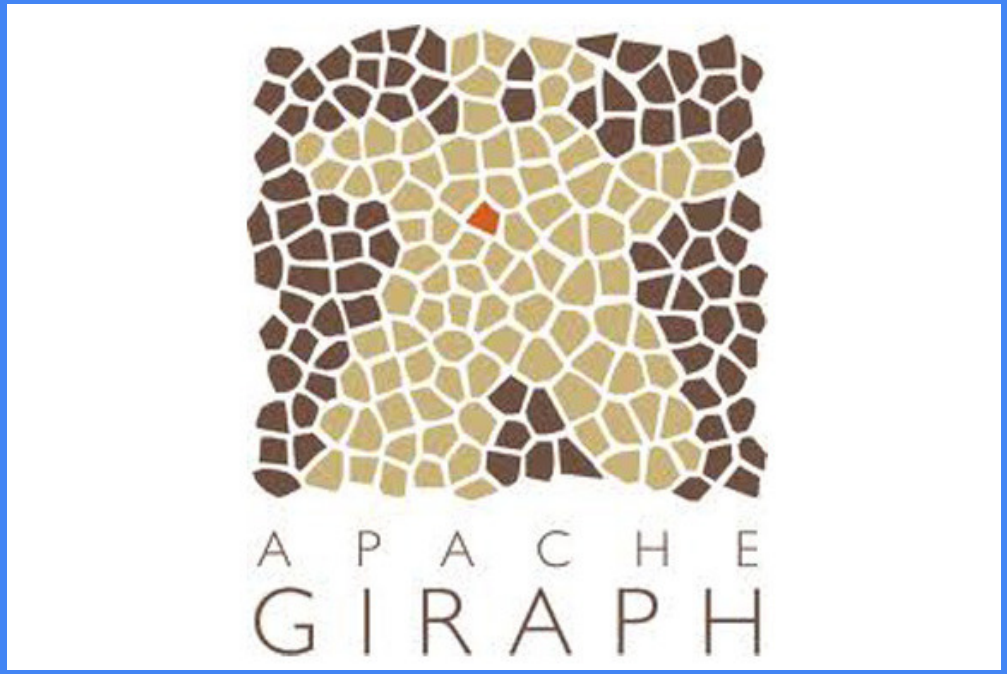
Batch Model



DISCO

HPCC SYSTEMS®

Graph Model



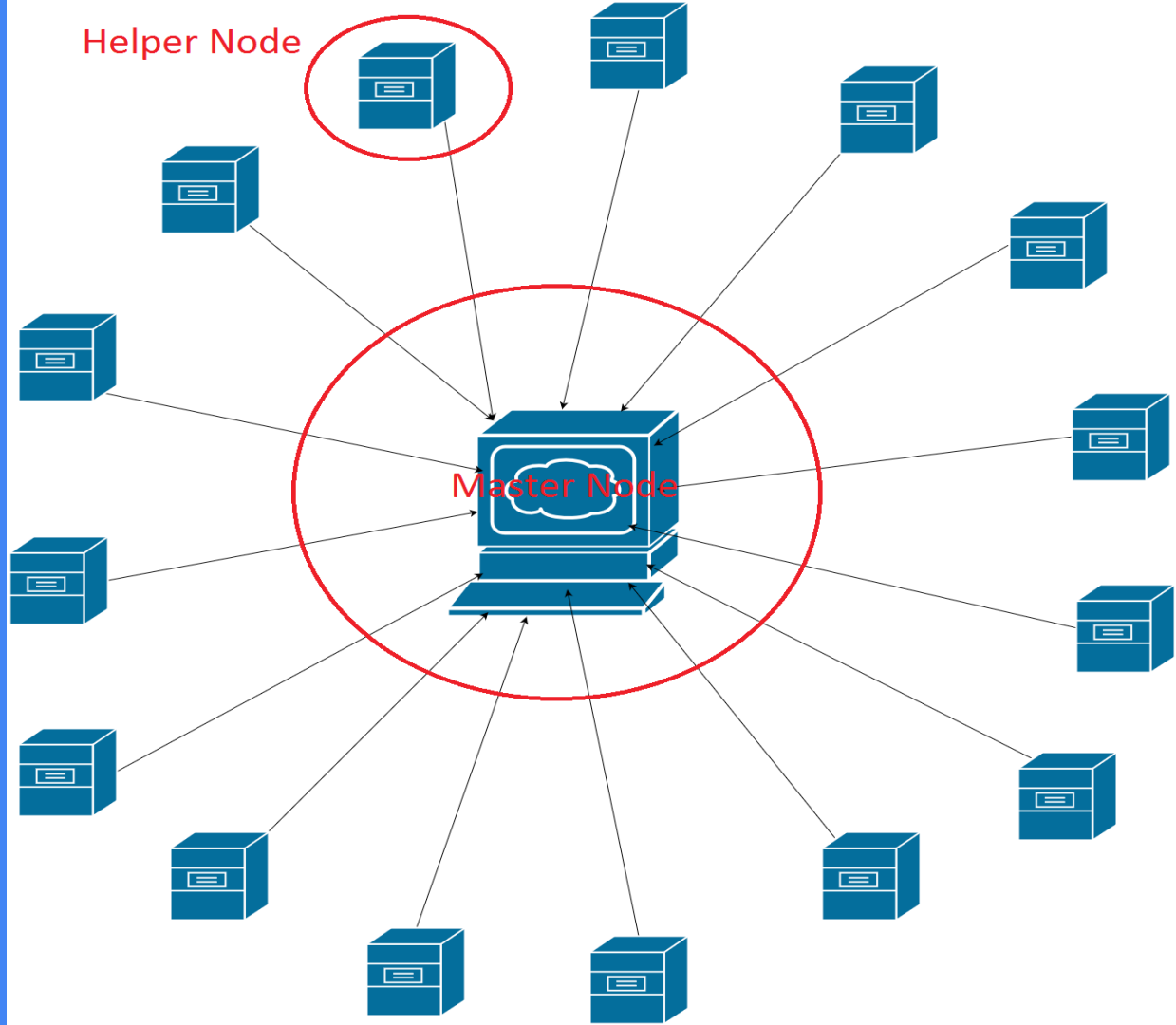
Streaming Model

samza

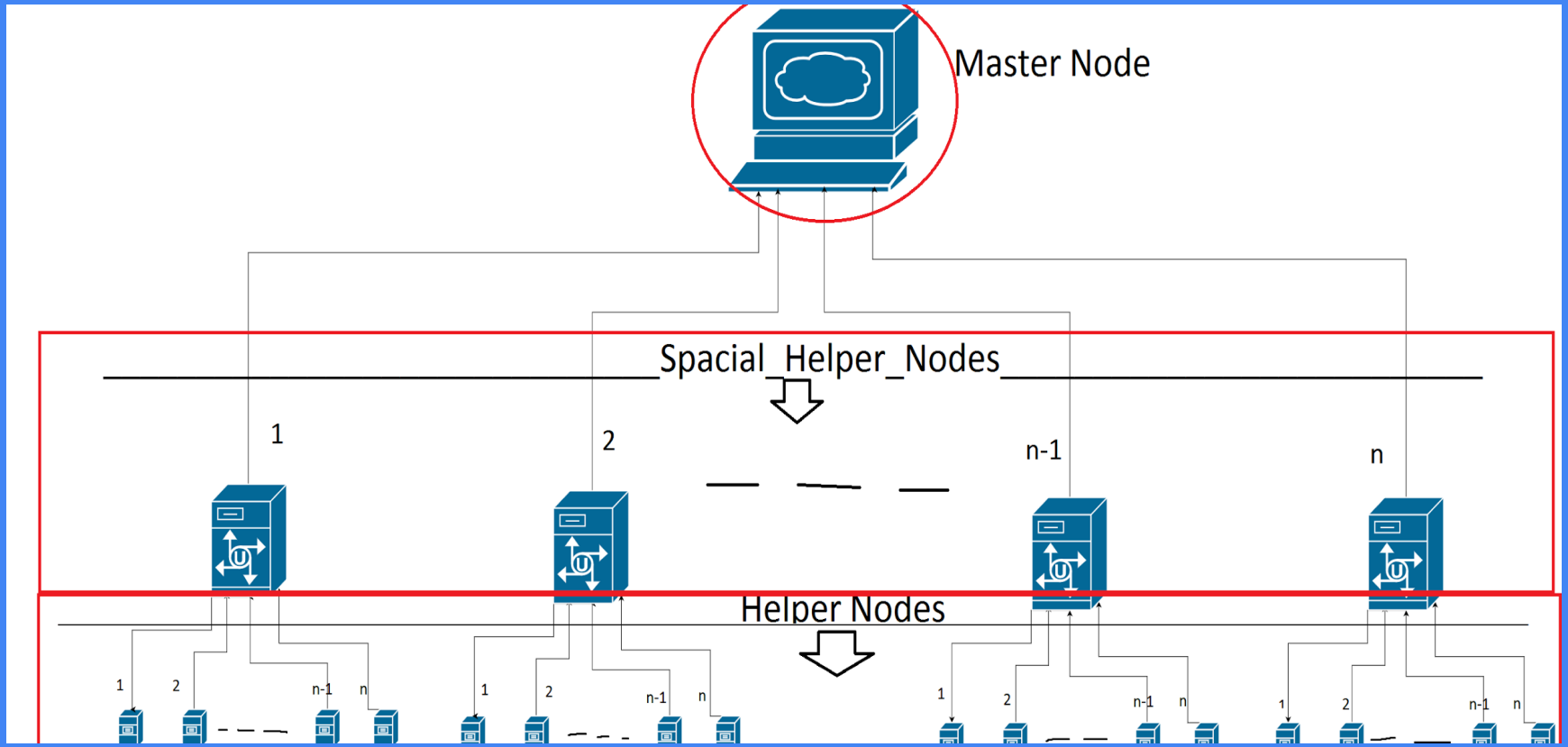
DataFlow Model

Problems

MapReduce at large Scale with commodity Hardware



Suggested Solution



Conclusion

What we have : lot's of data that is unique in one way or the other.

What we want : process that data in less time and less money.

What we have done : developed many simple data models to process complex real life data with our tools.

What we need : more data models and tools that can express more complex data and Hard Deadline Data processing system that guarantees job completion in fixed time

Questions ?

Thank you

Saumya Gupta
Surajnath Sidh

saumya.gupta@st.niituniversity.in
surajnath.sidh@st.niituniversity.in